U.PORTO

FEP FACULDADE DE ECONOMIA UNIVERSIDADE DO PORTO

Determinação do *Customer Lifetime Value*
Aplicação ao Retalho Alimentar

por

Patrícia Manuela Martins Castro

Dissertação de Mestrado em Modelação, Análise de Dados e Sistemas de Apoio à Decisão

Orientada por

Maria Paula de Pinho de Brito Duarte Silva

2017

Agradecimentos

Nesta fase tão importante na minha vida, não poderia deixar de agradecer a várias pessoas e entidades que contribuíram para a realização desta Dissertação.

À Professora Paula Brito pela excelente orientação, pela paciência e apoio.

À empresa que facultou a informação para a realização da dissertação.

A Deus por me orientar e ajudar nesta fase tão importante da minha vida.

A toda a minha equipa, a todas as meninas do clube marmita, à minha diretora Liliana Bernardino e, especialmente à minha chefe e grande amiga Ana Freitas que esteve sempre presente, obrigada pela compreensão, ajuda, disponibilidade e paciência. Muito obrigada mesmo.

Ao Carlos por me acompanhar nesta caminhada desde do início, pela paciência, pela ajuda e companheirismo.

À minha irmã Fátima por me apoiar nesta fase, pela paciência, pelo encorajamento e incentivo e por acreditar no meu trabalho.

À minha mãe pela melhor educação e por me ter ensinado a lutar pelos objetivos na vida e ao meu pai pela compreensão e apoio.

À Sílvia que sempre me encorajou, me animou nesta caminhada que também era dela, o meu sincero obrigada pelas suas palavras de coragem e orientação.

A todos que não se encontram aqui mencionados, mas que de certa forma contribuíram direta ou indiretamente para a realização desta dissertação.

O meu sincero OBRIGADA!

Resumo

Cada vez mais existe um olhar atento sobre o consumidor e o seu comportamento, pois dele depende a performance de qualquer empresa. Mais do que conhecer é necessário prever o seu comportamento no futuro para que com esse conhecimento sejam tomadas decisões de negócio.

O *Customer Lifetime Value* (CLV) é o valor que o cliente terá para a empresa a longo prazo, que no contexto da área em estudo é o valor de vendas liquidas reportadas que o cliente efetuará na empresa no ano seguinte.

Nesta dissertação foram desenvolvidos três modelos de previsão:

- Regressão Linear com variáveis de negócio que caracterizam o cliente segundo o seu comportamento de compra, envolvimento com a empresa e a informação sociodemográfica;
- Modelo de Pareto/NBD composto pelos dois submodelos de previsão Pareto/NBD e Gamma/Gamma;
- Regressão Linear com as variáveis de entrada definidas para aplicação do Modelo de Pareto/NBD.

A comparação entre estes três modelos foi realizada com recurso ao Erro Quadrático Médio (RMSE), ao Erro Absoluto médio (MAE) e o Coeficiente de Correlação de *Spearman* entre os valores reais e estimados.

O conhecimento do CLV permite à empresa ser mais assertiva no investimento dos clientes, mais eficaz na aplicação de estratégias de marketing e trabalhar melhor a fidelidade dos clientes.

Os resultados obtidos desta dissertação serão aplicados na empresa de retalho e irão sustentar a tomada de decisão nas estratégias do cliente.

Abstract

In the past years there has been an increase in studying customers and their behavior, since the performance of the companies relies on it.

More than understanding the customers, there is a need to predict their future behavior to make sustained strategic business decisions.

The Customer Lifetime Value (CLV) is the long-term value that a customer will bring to a company. In the field concerned by the present study, CLV is defined as the reported net sales value per customer on the following year.

In this dissertation three predictive models were developed:

- Linear regression with business variables that characterize the customer according to his shopping behavior, his engagement with the company and his sociodemographic data
- Pareto/NBD model that aggregates two Pareto/NBD and Gamma/Gamma sub-models
- Linear regression with the same variables as the ones used for the Pareto/NBD model.

The models' predictions were compared based on root mean square error (RMSE), the mean absolute error (MAE) and the Spearman correlation coefficient between observed and predicted values.

The Knowledge of the CLV allows the company to be more assertive on the investment made on customers, more effective when applying the marketing strategies and better promote customer's loyalty.

The models and results obtained in this dissertation will be applied on the retail company to support the customer strategic decision making.

Índice

Índice de tabelas

Índice de figuras

1 Introdução

O retalho alimentar e o comportamento do consumidor têm sofrido, ao longo dos últimos anos, uma grande evolução. As empresas de retalho têm crescido em número e a concorrência é cada vez maior, o que traz a necessidade de novas estratégias e definição de objetivos diferentes e ambiciosos. O comportamento do cliente também se alterou, pois com um leque de produtos e serviços disponíveis este tornou-se mais exigente e consciente do que gosta, do que é mais necessário, de qual é a proposta que trará mais benefícios, qual o produto/marca com maior qualidade entre uma série de outros requisitos. Consequentemente o cliente já não é tão fiel a nenhuma companhia, como acontecia no século XX, optando por fazer as compras no local que melhor respeite as suas necessidades no momento.

Assim, as empresas têm investido num maior conhecimento do seu negócio para que as suas decisões futuras sejam sustentadas em acontecimentos passados, o que consequentemente levará a uma decisão mais assertiva e eficaz.

A Análise de Dados é uma área com um desenvolvimento crescente nos últimos anos, propondo metodologias e modelos que dão resposta a vários problemas, em particular das empresas de retalho alimentar.

Um dos maiores problemas com que as empresas de retalho se deparam atualmente é a enorme necessidade de prever o que irá acontecer no futuro. A ideia deixou de ser trabalhar nos problemas após eles terem ocorrido, mas sim antes deles acontecerem. Para isso o negócio precisa de estar preparado de forma a permitir a tomada de decisões ou a aplicação de ações que irão barrar ou estimular um acontecimento, dependendo da estratégia. Por exemplo, quando é previsto que as vendas irão diminuir a empresa quererá impedir ou minimizar a probabilidade de ocorrência desse acontecimento. Quando, por outro lado, a empresa sabe que irá crescer em vendas, este conhecimento dará confiança à empresa para tomar decisões mais ambiciosas e assim ter uma maior probabilidade de obtenção de melhores resultados.

A empresa de retalho alimentar onde esta dissertação foi desenvolvida já tem a informação transacional bem estruturada e um nível avançado de estudo de dados e aplicação de metodologias que a Análise de Dados sugere.

Segundo Khajvand & Tarokh (2011), um dos mais importantes desafios nas organizações que dependem diretamente do cliente final é o conhecimento de cliente, a capacidade de perceber as diferenças entre eles e ordená-los segundo o que cada um representa para a empresa. O estudo e previsão do *Customer Lifetime Value* (CLV) é um tema muito abordado na literatura na área de Marketing e Economia como um conhecimento indispensável e com uma grande aplicação nas estratégias da empresa. O CLV, de um modo geral é o valor do cliente a longo prazo. No capítulo da Revisão de Literatura serão abordadas as diferentes definições e metodologias referidas no contexto de cada empresa.

O objetivo desta dissertação é estimar, para cada um dos clientes da empresa de retalho alimentar em estudo, um valor que indica se o cliente terá uma maior ou menor importância num determinado período de tempo no futuro. Este valor será calculado tendo por base variáveis que descrevem o comportamento do cliente no passado.

Os principais e mais diretos benefícios que este conhecimento dará à empresa serão uma maior assertividade, eficácia e um melhor tratamento da questão da fidelidade dos clientes. Assertividade nos investimentos feitos ao nível do cliente, uma melhor seleção da estratégia a aplicar em diferentes patamares do valor do cliente e consequentemente uma maior probabilidade de obtenção de resultados positivos. Maior eficácia na definição de ações promocionais ou outras ações direcionadas, com aplicação de filtros aos grupos de clientes com maior necessidade de acompanhamento e a consequente diminuição do valor investido na ação. No que diz respeito à fidelidade do cliente o problema central é o abandono. Um CLV inferior induz uma baixa fidelização do cliente com a companhia e por sua vez uma alta probabilidade de abandono. Assim, estes clientes necessitam de mais atenção e incentivos de forma a estimular o aumento do número de visitas ou valor por visita às lojas da empresa.

1.1 Definição do *Customer Lifetime Value*

O *Customer Lifetime Value* assume diferentes definições na literatura conforme o contexto onde este está a ser desenvolvido e a visão de negócio de quem o desenvolve. O contexto onde esta dissertação será feita é uma empresa de retalho alimentar onde existe um programa de fidelização de cliente há cerca de 10 anos. Uma vez que o cliente não tem nenhum contrato com a empresa, o contexto é denominado como não-contratual (Aghaie, 2009).

Para a empresa de retalho alimentar em estudo nesta dissertação, a definição do CLV que se considera mais adequada é:

O valor de vendas líquidas reportadas que o cliente gastará na empresa num dado período de tempo futuro.

As vendas líquidas reportadas são calculadas a partir da subtração dos descontos líquidos às vendas líquidas. As vendas líquidas resultam da subtração entre as vendas brutas e o valor do IVA praticado aquando da transação. O desconto líquido é o valor que o cliente acumula em cada compra sem o valor do IVA, que pode posteriormente ser utilizado numa transação futura.

Inicialmente o valor do cliente seria o retorno do investimento que este daria à empresa num determinado período de tempo, mas dado a inacessibilidade do custo por cliente, a sua definição foi reformulada.

1.2 Metodologia

Dada a diversidade de metodologias propostas na literatura, o cálculo do CLV nesta dissertação foi efetuado através de três metodologias diferentes: a Regressão Linear Múltipla com variáveis de negócio (Glady, Baesens, & Croux, 2009; Singh & Jain, 2013), o Modelo de Pareto/NBD (*Negative Binomial Distribution*) (Fader, Hardie, & Lee, 2005a; Fader, Hardie, & Lee, 2005b ; Glady *et al*., 2009; Singh & Jain, 2013) e a Regressão Linear Múltipla com as variáveis de entrada do Modelo de Pareto/NBD.

A Regressão Linear é um método de previsão de uma variável resposta a partir de um conjunto de variáveis explicativas independentes, cada uma com um coeficiente que define a relação entre a variável explicativa e a variável resposta. As variáveis explicativas utilizadas são variáveis que descrevem o comportamento de compra do cliente na empresa. O modelo de regressão é um modelo usual dentro da Análise Preditiva, aplicado a previsões de variáveis continuas e de fácil implementação.

O Modelo de Pareto/NBD é uma metodologia abordada pela primeira vez em 1987 por David C. Schmittlein (Schmittlein, Morrison, & Colombo, 1987), que inicialmente baseava-se na previsão do número de transações e na probabilidade do cliente estar ativo, mas subsequentes trabalhos foram desenvolvidos de versões modificadas do modelo original, com melhorias do já existente e/ou trabalhos que complementaram e enriqueceram o resultado deste método. Destes diversos trabalhos nasceu o Submodelo Gamma/Gamma (Fader, Hardie, & Lee, 2005b), que prevê o gasto médio de cada cliente em cada transação. O CLV é obtido através da multiplicação do número de transações estimado pelo gasto médio também previsto.

Comparação de Metodologias

Os resultados serão avaliados por recurso a três medidas de precisão dos modelos, o Erro Quadrático Médio (RMSE), Erro Absoluto Médio (MAE) e o Coeficiente de Correlação de *Spearman* (Glady *et al.*, 2009). Glady e os restantes autores calcularam o erro apenas com 99% das melhores previsões, ignorando os clientes que têm um comportamento muito atípico, argumentando que esta metodologia evita que estes clientes dominem os resultados da análise.

O RMSE é definido então como

1.1 Erro Quadrático Médio

$$RMSE = \sqrt{\frac{1}{0.99 \times n} \sum_{i \in BP} (\widehat{CLV}_i - CLV_i)^2}$$

A medida MAE é calculada através

1.2 Erro Absoluto Médio

$$MAE = \frac{1}{0.99 \times n} \sum_{i \in BP} |\widehat{CLV}_i - CLV_i|$$

com n sendo o número de observações em análise e *BP* as melhores previsões obtidas através do modelo. Valores baixos de RMSE e MAE significam que a distância entre os valores estimados e os valores reais do CLV é pequena, ou seja, indiciam uma boa previsão. O Coeficiente de Correlação de *Spearman* mede a correlação entre o *rank* dos valores estimados e o *rank* dos valores observados do CLV. Um coeficiente de correlação acima de 0,5 indicia uma boa previsão do valor do CLV. Com base nestas medidas seleciona-se a melhor metodologia.

Estrutura da Dissertação

Esta dissertação está divida em 6 capítulos que separam as diferentes fases do estudo e desenvolvimento da previsão do CLV. No início de cada capítulo encontra-se um pequeno resumo dos temas abordados que estão também eles separados em subcapítulos.

No capítulo 1 é uma introdução ao tema, com o esclarecimento da motivação do desenvolvimento desta dissertação e a sua importância na área de retalho, a definição do CLV neste contexto e metodologia utilizada na previsão do mesmo.

O capítulo 2 descreve a literatura existente sobre o CLV, a importância do conhecimento do mesmo, as diferentes definições para cada contexto de negócio, as metodologias aplicadas nesses diferentes contextos e uma descrição mais aprofundada das metodologias aplicadas nesta dissertação.

Antes do desenvolvimento dos diferentes modelos fez-se uma descrição extensa da base de dados utilizada, onde são referidos alguns termos específicos da área de retalho alimentar e a importância de cada variável na previsão do CLV. Acrescente disso também foi feita uma análise exploratória e o tratamento necessário de algumas variáveis. Todo este estudo da base de dados encontra-se no capítulo 3.

No capítulo 4 contem a descrição de todas as metodologias aplicadas para a previsão do CLV.

A discussão de resultados é descrita no capítulo 5, com a comparação dos resultados obtidos das três metodologias.

No capítulo 6 são apresentadas as principais conclusões, as limitações e considerações finais relevantes, bem como algumas recomendações para trabalho futuro.

2 Revisão da Literatura

"Os últimos anos têm assistido a uma explosão de pesquisas em *Customer Lifetime Value"* (Singh & Jain, 2013). Como se define, como se calcula, o que o influencia e o porquê de ser tão importante são algumas das questões que autores como Gupta, Pfeifer, Singh, Fader e Glady procuram responder. Os anteriores estudos sobre cliente procuravam conhecê-lo, compreendê-lo e classificá-lo segundo o seu comportamento no passado. Atualmente, e paralelamente à evolução ocorrida no mercado de retalho e na área de Análise de Dados, a visão sobre o cliente também evoluiu e os gestores reconheceram a importância vital da satisfação e lealdade dos atuais clientes e vêm-no como o principal contributo para uma vantagem em relação aos concorrentes. Assim, mudaram a sua estratégia de marketing para uma versão onde o cliente está no centro da estratégia – *Customer-Centric* (Aghaie, 2009).

Desta forma, foram criadas novas áreas e espaços dedicados à informação de cliente para que esta seja sempre viável, atualizada e esteja apta para ser utilizada nas diversas ações direcionadas. Exemplos deste facto são o *Knowledge Management* (KM) e *Customer Relationship Management* (CRM) (Aghaie, 2009) que são orientações de como devemos tratar os dados gerados sucessivamente e transformá-los em conhecimento de negócio e numa importante base para o processo de decisões centradas no cliente.

2.1 Importância do conhecimento do valor do cliente

"Estudos indicam que a aquisição de um novo cliente custa cinco vezes mais do que a retenção de um cliente existente e que este último é muito mais rentável" (Aghaie, 2009)

Na mesma linha de raciocínio, Gupta, Lehmann e Stuart (2004) referem que as empresas, os empresários e analistas estão demasiado focados em vendas e investimentos porque acreditam que há um impacto direto no valor da empresa. Mas a conclusão retirada do estudo é que a taxa de retenção tem um impacto significativamente maior no valor da empresa, impondo que os empresários devem prestar uma maior atenção a esta variável. Os autores referem ainda que o CLV é considerado um ativo para a empresa contribuindo

para o valor da mesma. E desta forma sublinham mais uma vez a necessidade de ser avaliado, como acontece com outros ativos.

Por todas estas razões é importante comunicar corretamente com os clientes já existentes na empresa, evitando falhas e a consequente insatisfação do mesmo. Para que seja possível dar resposta às necessidades do cliente temos de o conhecer e saber como incentivar a sua compra.

Dependentemente da área de negócio ou estratégia onde o CLV será aplicado existem vários exemplos da sua aplicação na literatura. Segundo Glady (2009) o valor de um cliente é importante para a deteção dos clientes mais e menos valiosos para depois a empresa definir a quais deve prestar maior e menor atenção. Ainda Khajvand & Tarokh (2011) referem que o CLV é mais fácil de ser trabalhado e usado na tomada de decisões de marketing se for calculado em segmentos de clientes e sugere também que o CRM poderia ser adaptado às características de cada segmento. Nas ações direcionadas, o CLV tem também um papel importante pois o orçamento pode ser mais eficiente quando o valor do grupo de clientes a ser selecionado é conhecido (Glady *et al.*, 2009).

2.2 Definição do CLV

A definição do CLV é muito ambígua chegando mesmo a ser, em algumas situações, contraditória. Isto acontece porque o CLV não tem uma definição única, o seu significado varia conforme o ambiente onde ele está a ser calculado. Por esta razão, para a definição do CLV e mesmo das variáveis que irão defini-lo é necessário conhecer o contexto do negócio onde ele vai ser calculado (Singh & Jain, 2013). Assim foram criadas classificações que caracterizam cada contexto de negócio segundo o tipo de contrato (contratual e não contratual), o tempo ocorrido entre interações do cliente com a empresa (contínuo ou discreto) e o Gasto Médio por cliente (fixo ou variável) (Singh & Jain, 2013). Várias referências apontam que um dos maiores problemas na previsão do valor do cliente é o desconhecimento do momento em que o cliente abandona a empresa (Fader et al., 2005b; Singh & Jain, 2013). Esta classificação de contextos não é só importante na organização das definições do CLV, mas também na seleção de metodologias adequadas a cada uma das classes.

O mercado de retalho alimentar é um contexto não contratual pois não é possível saber o momento em que o cliente abandona. O seu tempo entre compras e o gasto em cada compra é muito variável o que dificulta a previsão do seu abandono.

Os artigos mais citados sobre CLV (Gupta *et al.*, 2004; Pfeifer, Haskins, & Conroy, 2004) defendem que o CLV nada tem a ver com a rentabilidade do cliente, definindo-o como o valor presente dos fluxos de caixa futuros atribuídos à relação com o cliente. Por outro lado, outros significados de CLV têm por base exatamente o lucro que o cliente dá à empresa, onde o define como o valor descontado dos lucros futuros (Glady *et al.*, 2009). Aghaie (2009) em concordância com Glady, define o CLV como um conjunto de dois comportamentos e/ou variáveis: o lucro do cliente para a empresa e o seu contributo nos efeitos 'boca-a-boca' ou em redes dentro de comunidades de clientes. Esta variável que mede a reputação que cada empresa tem, ao nível do cliente, não é fácil de obter. Apenas empresas com acesso privilegiado ou empresas online é que conseguem analisar por cliente esta variável, como é o caso analisado por Aghaie (2009).

Numa abordagem algo diferente das definições já referidas, temos a definição de Fader (2005b) e Khajvand e Tarokh (2011), onde referem que o CLV é simplesmente o valor do cliente na segmentação RFM[1] (*Recency, Frequency and Monetary*) no futuro.

Tal como foi referido acima, nenhuma destas definições está errada, pois estão adequadas a cada um dos contextos onde se encontram inseridas. É necessário perceber as razões que estão na base de cada uma delas para que depois, no seu próprio contexto de negócio, seja mais fácil de definir e calcular o seu CLV.

2.3 Metodologia

A metodologia a ser utilizada nesta dissertação tem por base o trabalho de Glady (2009), onde é desenvolvida uma nova versão do modelo de Pareto/NBD e se compara com o Modelo de Pareto/NBD e a Regressão Linear, tendo por base a informação de serviços

[1] RFM é uma metodologia de segmentação de clientes tendo por base três variáveis: *Recency* (quanto tempo desde a última compra), *Frequency* (o número de visitas à loja) e *Monetary* (valor gasto pelo cliente na empresa).

prestados de um banco. Nesta dissertação será desenvolvido o Modelo antigo de Pareto/NBD (Schmittlein et al., 1987).

Outros trabalhos como Aghaie (2009), Fader et al (2005b), Khajvand & Tarokh (2011), definem o valor do cliente a partir da segmentação RFM e posteriormente aplicam métodos de previsão como Redes Neuronais *Feedforward* Multicamadas (MFNN) ou ARIMA de forma a prever o segmento do cliente no futuro.

No artigo de Singh & Jain (2013), onde é desenvolvida a classificação de contextos referida anteriormente, os autores sugerem várias metodologias em cada uma das classes. Dentro dos contextos contratuais, onde é conhecida a data de abandono do cliente, os autores sugerem Modelo Estrutural Básico do CLV, Modelos de Regressão/RFM e Modelos de Taxa de Risco. Para os contextos não contratuais a literatura sugere Pareto/NBD (utilizada nesta dissertação), Beta-Geométrica/NBD, Modelos de Cadeias de *Markov* e Cadeias de *Markov* Monte Carlo.

A Regressão, apesar de estar nas sugestões de contextos contratuais, têm uma grande aplicabilidade também em contextos não contratuais onde se conseguem resultados igualmente bons (Singh & Jain, 2013).

Dado o contexto onde esta dissertação se insere e o conhecimento já existente na empresa optou-se pelas duas metodologias, Regressão Múltipla e Pareto/NBD, e uma posterior seleção da que tiver melhor desempenho, pois são metodologias que se adaptam à realidade da empresa em estudo e permitem a fácil aplicabilidade no quotidiano da mesma. A Regressão Linear Múltipla será utilizada no desenvolvimento de dois modelos com variáveis explicativas distintas: um composto pelas variáveis de negócio, outro pelas as variáveis de entrada do Modelo Pareto/NBD (maior detalhe no subcapítulo 2.3.2).

2.3.1 Regressão Linear Múltipla

A Regressão Linear Múltipla é um método de previsão muito aplicado e extensamente estudado em Análise de Dados e Estatística. Este método consiste em modelar uma determinada variável aleatória (variável resposta) segundo uma ou mais variáveis independentes (variáveis explicativas). Quando o modelo apenas utiliza uma variável independente para a previsão da variável resposta o modelo designa-se por Regressão

Linear Simples. O modelo a ser aplicado nesta dissertação tem em consideração mais do que uma variável explicativa e por essa razão é mais completo do que a Regressão Linear Simples.

Nesta abordagem o cálculo do CLV é calculado através da expressão:

2.1 Modelo de Regressão Linear Múltipla

$$\widehat{CLV}_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i,$$

onde para cada cliente i, $X_{i1}, X_{i2}, \ldots, X_{ip}$ são as p variáveis explicativas do CLV; β_0 é a ordenada na origem, ou seja, o valor do CLV quando todas as variáveis explicativas (X_{ip}) são 0; $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes que representam os declives parciais, isto é, quanto é que a variável resposta varia por uma unidade da variável explicativa; e ε_i é o erro de previsão associado ao CLV do cliente i.

Este modelo tem alguns pressupostos que necessitam de ser cumpridos para que a veracidade e precisão dos resultados sejam garantidas, tais como:

1. Os valores de ε_i seguem uma distribuição Normal, com média 0 e variância constante σ^2;
2. ε_i's são variáveis aleatórias independentes;
3. As variáveis explicativas não devem ser correlacionadas.

O valor que o cliente terá no mercado de retalho alimentar é influenciado por variáveis de comportamento de compra habitual e variáveis que medem o envolvimento do cliente com a companhia. A Regressão Linear Múltipla é muito flexível em relação à inclusão de novas variáveis, o que torna este método muito vantajoso na adaptação a diferentes contextos e áreas de negócio.

A previsão de valores no futuro, na maioria das áreas de negócio, depara-se com um problema que complica bastante a obtenção de boas previsões: a sazonalidade. No mercado de retalho alimentar especificamente, o período sazonal é de 12 meses, pois existem eventos ao longo do ano onde os consumidores alteram o seu comportamento de compra tais como o Natal, a Páscoa e as férias de Verão. O Modelo de Regressão Múltipla não leva em consideração a sazonalidade dos dados e por esta razão só é possível efetuar previsões entre intervalos anuais. Esta é uma das limitações deste modelo. Outras

limitações são referidas na literatura, como por exemplo a necessidade de dividir a amostra em duas partes: uma parte ser o conjunto de dados de treino do modelo e a outra para o teste do modelo (Glady et al., 2009).

2.3.2 Modelo de Pareto/NBD

O Modelo de Pareto/NBD é um modelo mais complexo que a Regressão Linear Múltipla. Esta metodologia compreende dois modelos de previsão: o Submodelo Pareto/NBD e o Submodelo Gamma/Gamma.

O submodelo Pareto/NBD estima o número de transações $\hat{x}_{i,T_i+k}$ que um cliente i fará num dado período $T_i + k$. Este modelo necessita de quatro variáveis: o tempo T_i (*cohort*) desde o momento de entrada do cliente na empresa até "agora"; a frequência, o número de vezes que o cliente visitou a empresa desde a sua entrada até agora representado por $x_i \equiv x_{i,T_i}$, sendo que o total de transações será $x_i + 1$ porque a primeira transação é onde o cliente se torna ativo; e por fim o tempo entre a data da primeira e da última compra do cliente na empresa, *recency*[2], t_i. O Submodelo Gamma/Gamma estima o lucro por transação para cada cliente i no período $T_i + k$, representado por $\hat{m}_{i,T_i+k}$.

Tal como a Regressão Linear Múltipla, o Modelo de Pareto/NBD também tem requisitos que a base de dados deve cumprir para que seja aplicável. Assim os pressupostos do Modelo de Pareto/NBD (Glady *et al.*, 2009; Singh & Jain, 2013) são:

1. Enquanto ativo, o cliente i faz compras segundo um processo de *Poisson* com taxa λ_i;
2. Cada cliente permanece ativo durante um tempo que é exponencialmente distribuído com uma taxa de abandono de μ_i;
3. A taxa de compra λ_i para diferentes clientes é distribuída de acordo com a Distribuição *Gamma* para o total da população, com parâmetro r, com $\alpha > 0$. A taxa de compra média para todos os clientes é $E[\lambda] = {}^{r}/_{\alpha}$ e variância é ${}^{r}/_{\alpha^2}$;

[2] É importante referir que o conceito *recency* no Modelo do Pareto/NBD é diferente do conceito de *recency* na segmentação RFM referida anteriormente.

4. As taxas de abandono μ_i são distribuídas de acordo com uma distribuição de *Gamma* diferente para todos os clientes. A taxa de abandono média para todos os clientes é $E[\mu] = {}^{s}/_{\beta}$ e variância é ${}^{s}/_{\beta^2}$;
5. A taxa de compra λ_i e a taxa de abandono μ_i são independentemente distribuídas;
6. Para cada cliente i o lucro por transação é independente do número de transações;
7. O valor do lucro por transação segue uma Distribuição *Gamma* com parâmetro de forma px_i e parâmetro de escala ${}^{1}/_{\nu_i}$;
8. Os valores de ν_i seguem uma distribuição Gamma com parâmetro de forma q e parâmetro de escala ${}^{1}/_{\gamma}$;

Os parâmetros $r, \alpha, s, \beta, p, q, \gamma$ serão estimados através do Método de Máxima Verosimilhança (MLE).

Assim o CLV é obtido através da expressão:

2.2 Modelo de Pareto/NBD

$$\widehat{CLV}_{i,h} = \sum_{k=1}^{h} \frac{(\hat{x}_{i,T_i+k} - \hat{x}_{i,T_i+k-1})\widehat{m}_i}{(1+d)^k}$$

Para o cliente i, o CLV é calculado para um horizonte de h períodos onde $\hat{x}_{i,T_i+k}$ é o número estimado de transações no período $T_i + k$, $\widehat{m}_{i,T_i+k}$ é o valor estimado do lucro dado pelo cliente à companhia em cada transação e d é a taxa de desconto. Segundo Glady (2009) a taxa de desconto é uma constante; no exemplo onde a metodologia foi aplicada esta taxa de desconto correspondia ao Custo do Capital Médio Ponderado da empresa. Para o contexto de empresa em estudo, o valor da taxa de desconto é desconhecido, por isso não será incorporado no cálculo do CLV.

2.3 Modelo de Pareto/NBD aplicado ao contexto em estudo

$$\widehat{CLV}_{i,h} = \sum_{k=1}^{h} \hat{x}_i \times \widehat{m}_i$$

O Modelo de Pareto/NBD é um modelo menos flexível do que a Regressão Linear Múltipla, mas apresenta outras vantagens tais como a não necessidade de divisão de amostra (quando prevemos através da Regressão Linear Múltipla são necessárias duas

amostras: uma amostra onde se irá basear a modelação e a outra para validação da previsão) e a possibilidade de o período de estimação de valores ser inferior ao período de sazonalidade.

3 Base de Dados

A informação trabalhada foi fornecida pela empresa de retalho alimentar e trata-se da informação transacional e sociodemográfica dos clientes dessa mesma empresa. O período de informação utilizado foram dois anos, de janeiro 2015 a dezembro 2016.

Na base de dados o cliente está identificado no campo *id* através de um código irreal. É importante referir que as variáveis apresentam valores mascarados de modo a garantir a confidencialidade dos valores reais da empresa.

Na previsão do CLV trabalhou-se com dois conjuntos de variáveis, um conjunto por tipo de metodologia. Como já foi referido anteriormente, o Modelo de Pareto/NBD tem três variáveis de entrada já definidas (número de transações repetidas, dias entre a primeira e última compra e dias desde do primeiro dia de compra até ao final do período de análise). Já na Regressão Linear as variáveis de entrada ou variáveis explicativas podem ser todas as que o utilizador achar importantes na previsão do CLV.

Deste modo, o conjunto de variáveis selecionadas para estimação do CLV de cada cliente na Regressão Linear descrevem-no segundo o seu comportamento de compra, o seu envolvimento/fidelidade com a empresa e a sua informação sociodemográfica.

Relativamente às variáveis de comportamento de compra foram consideradas:

1. O número de transações, isto é, o número de vezes que o cliente efetua compras em alguma das lojas da empresa no período considerado. É uma variável quantitativa discreta, com valores positivos.
2. Gasto bruto do cliente, que corresponde ao valor de vendas brutas totais do cliente, no período, nas lojas da empresa. Variável quantitativa contínua com valores positivos.
3. Gasto líquido do cliente que corresponde ao valor de vendas líquidas totais do cliente, no período, nas lojas da empresa no período considerado. Variável quantitativa contínua com valores positivos.
4. Gasto líquido reportado (VLR) do cliente que corresponde ao valor de vendas líquidas totais menos o valor de desconto líquido do cliente no período nas lojas da empresa. Variável quantitativa contínua com valores positivos.

5. Valor de VLR por transação, designada por Cesta Média Reportada. Este valor é obtido através da divisão das VLR pelo número de transações do cliente no período em análise. É uma variável quantitativa contínua, com valores positivos.
6. Percentual de semanas com compra no período, que resulta da divisão do número de semanas onde o cliente tem pelo menos uma compra pelo total das semanas no período. É uma variável quantitativa contínua onde os seus valores variam entre 0 (exclusive) e 1.
7. Número de dias desde a última compra até ao último dia do período em análise (*Recency*). Esta variável permite saber o quão recente é a última compra do cliente no período. Variável quantitativa discreta que toma valores positivos.
8. Valor de desconto bruto acumulado no cartão de fidelização durante o período em análise. É uma variável quantitativa contínua com valores positivos.
9. Valor de desconto líquido direto no período de análise. Desconto líquido deduzido diretamente no valor a pagar pelo cliente. É uma variável quantitativa contínua com valores positivos.
10. Valor de desconto total líquido, que soma o desconto líquido acumulado em cartão com o desconto líquido descontado diretamente no valor a pagar pelo cliente. Trata-se de uma variável quantitativa contínua de valores positivos.
11. Segmento Valor da Segmentação Valor da empresa[3]. Corresponde ao segmento que é a moda dos segmentos mensais desse cliente. Variável qualitativa com 4 categorias: *Loyal, Frequent, Occasional* e Sem Valor.

Relativamente às variáveis que medem o envolvimento do cliente serão consideradas:

12. Ter ou não ter cartão de crédito com vantagens adicionais ao cliente em compras no ecossistema[4] da empresa em estudo. É uma variável binária com valores {0 – não ter,1 – ter}.
13. Ter ou não ter um desconto adicional direcionado para um grupo de clientes segundo um determinado critério. É uma variável binária com valores {0 – não ter,1 – ter}.

[3] A Segmentação Valor tem por base a metodologia da segmentação RFM (*Recency, Frequency* e *Monetary*) mas considerando apenas duas das três variáveis naturais dessa segmentação (*Frequency* e *Monetary*)

[4] A empresa de retalho em estudo tem parceria com um conjunto de empresas de diversas áreas diferentes do retalho alimentar: quiosques, parafarmácias, vestuário, *petcare* entre outros.

14. Percentagem de vendas brutas dos produtos do não-alimentar no gasto total do cliente no período. Uma vez que se está a analisar uma empresa de retalho alimentar, onde o seu *core* são produtos alimentares, os clientes habituais de produtos não-alimentares são clientes mais envolvidos com a marca do que os clientes que apenas compram produtos alimentares. É uma variável quantitativa contínua cujos valores variam entre 0 e 1.
15. Vendas brutas do não-alimentar. Variável quantitativa contínua de valores positivos.
16. Percentagem de vendas brutas nas lojas do ecossistema (exceto lojas de retalho alimentar) sobre o gasto total do cliente no ecossistema. Nesta dissertação o foco são as lojas de retalho alimentar, mas se os clientes destas mesmas lojas efetuarem também compras no ecossistema da empresa este será um cliente mais fiel do que os clientes sem compras nas restantes lojas do ecossistema. Esta é uma variável quantitativa contínua cujos valores variam entre 0 e 1.
17. Vendas brutas nas lojas do ecossistema (exceto lojas de retalho alimentar). Variável quantitativa contínua de valores positivos.
18. Número de parceiros do ecossistema com pelo menos uma compra no período de análise. Variável quantitativa discreta de valores positivos.
19. Idade da conta do cliente (*lifetime*), em dias. Tempo decorrido desde da ativação do cartão até ao último dia do período em análise. Um cliente que aderiu ao cartão de fidelização recentemente não é tão fiel como o cliente que aderiu ao cartão desde a sua existência. Variável quantitativa discreta de valores positivos.

Relativamente às variáveis que caracterizam o cliente serão ainda consideradas:

20. Ser ou não ser segmento comercial na Segmentação *Grocer*[5], onde o segmento comerciante significa ter um grande valor para a empresa. É uma variável binária com valores {0 – não ser comerciante,1 – ser comerciante}.
21. Ter ou não ter filhos, através da Segmentação *Baby&Junior* existente na empresa. Se um cliente tem filhos terá um maior gasto no retalho alimentar. É uma variável binária com valores {0 – não ter,1 – ter}.

[5] Segmentação que classifica o cliente segundo o seu comportamento comercial e/ou abusivo. Esta segmentação contem os segmentos: Comerciais e Caça-Promoções.

22. Variável que sinaliza se o cliente está mais próximo de um ponto de venda da empresa ou de um da sua concorrência. Esta variável é binária com os seguintes valores {1 mais perto do parceiro / 0 mais perto da concorrência}
23. Densidade de lojas do retalho alimentar próximas da residência do cliente, é uma medida da quantidade de lojas da área de negócio em estudo. Quanto mais próximo de 0 maior a densidade (nº de lojas existentes na envolvente do cliente) e quanto maior, menor essa mesma densidade. Esta variável é quantitativa contínua positiva.
24. Número de membros do agregado familiar do cliente. Assume-se que quanto maior o agregado familiar maior será o gasto do cliente no retalho alimentar. Variável quantitativa discreta de valores positivos.
25. Idade do cliente, em número de dias desde o seu nascimento até ao último dia do período de análise. Variável quantitativa discreta de valores positivos.

A Tabela 1 apresenta um resumo das variáveis utilizadas para ser mais fácil para o leitor consultar o significado de cada uma e o formato utilizado.

Tabela 1 Quadro resumo das variáveis utilizadas no modelo

N	Nome da variável	Descrição	Tipo de Variável
Variáveis para Regressão Linear			
1	agregado	Número de elementos do agregado familiar do cliente	Quantitativa discreta
2	cart_uni	Ter ou não ter o cartão de crédito com benefícios adicionais	Binária
3	cesta_rep	Valor médio gasto por transação	Quantitativa continua
4	descontos	Valor descontado indiretamente em 2015	Quantitativa continua
5	dl_sp	Desconto líquido direto	Quantitativa continua
6	dl_tot	Desconto líquido total	Quantitativa continua
7	flag_closer_partner	Estar ou não estar mais perto de uma loja interna relativamente à concorrência	Binária
8	idade	Idade do cliente	Quantitativa discreta
9	lifetime	Idade da conta do cliente	Quantitativa discreta
10	nr_ins	Número de parceiros visitados pelo cliente em 2015	Quantitativa discreta
11	nr_trx	Número de transações em 2015	Quantitativa discreta
12	overall_stores_density	Densidade de lojas do retalho alimentar próximas da residência	Quantitativa continua
13	perc_sem	Percentagem de semanas com compras em 2015	Quantitativa continua
14	perc_vb_eco	Percentagem de vendas brutas despendidas nas lojas do ecossistema da empresa	Quantitativa continua
15	perc_vb_na	Percentagem de vendas brutas despendidas no não alimentar	Quantitativa continua
16	primav	Com ou sem desconto adicional	Binária

17	segm_baby_junior	Ter ou não ter filhos	Binária
18	segm_grocer	Ser ou não ser comerciante	Binária
19	segm_valor	Segmento Valor	Qualitativa
20	t_recency	*Recency* em 2015	Quantitativa discreta
21	vb	Vendas Brutas em 2015	Quantitativa continua
22	vb_eco	Vendas brutas despendidas nas lojas do ecossistema da empresa	Quantitativa continua
23	vb_nalim	Vendas brutas despendidas no não alimentar	Quantitativa continua
24	vl	Vendas Líquidas em 2015	Quantitativa continua
25	vlr	Vendas líquidas sem descontos em 2015	Quantitativa continua
26	vlr_resp	Vendas líquidas sem descontos em 2016 (variável resposta)	Quantitativa continua
Variáveis para Pareto/NBD			
1	x	Número de transações repetidas	Quantitativa continua
2	t	Número de dias entre o primeiro e o último dia de compra	Quantitativa discreta
3	T.cohort	Número de dias entre o primeiro dia de compra e o último dia de análise	Quantitativa discreta

3.1.1 Análise Exploratória dos Dados

A análise exploratória é uma das etapas mais importantes de qualquer projeto que tem por base informação. Só conhecendo bem as variáveis e os seus valores é que podemos ter noção das suas fragilidades e mais valias, permitindo também fazer as transformações adequadas para aplicação do método utilizado.

Os *softwares* utilizados para o tratamento dos dados e a aplicação das metodologias foram o SAS e o *software* R com o *package* BTYD (McCarthy & Wadsworth, 2014). O SAS na Regressão Linear Múltipla e no estudo exploratório das variáveis e o R na aplicação do Modelo do Pareto/NBD.

Na análise exploratória é importante analisar duas vertentes: como a variável se comporta individualmente e como se comporta em relação a outras variáveis. Assim é necessário fazer uma análise univariada e bivariada.

Análise Univariada

Na análise univariada é necessário medir a centralidade, dispersão e distribuição dos valores observados. De forma a medir a centralidade dos dados calcularam-se a média, a moda, a mediana e os quartis. Para avaliar a dispersão utilizaram-se as medidas mais conhecidas como o mínimo, o máximo, o desvio padrão e o coeficiente de variação. Para analisar a distribuição de cada variável recorreu-se à representação visual da distribuição empírica dos dados – o histograma.

Nas Tabela 2 e Tabela 3 registam os valores das principais medidas de centralidade e dispersão, assim como o número total de observações e o número de valores omissos por variável quantitativa.

Tabela 2 Principais medidas de centralidade e dispersão

Variable	Mean	Std Dev	Minimum	Maximum	N	N Miss
VLR	1124.9362	1421.0530	0.0107	94170.0141	3428972	0
VB	1387.4444	1735.9928	0.1188	116966.2358	3428972	0
VL	1208.8533	1519.1984	0.1007	95127.2813	3428972	0
DESCONTOS	96.4230	140.2641	0.0000	23979.4859	3428972	0
t_recency	40.0355	71.3083	0.0000	363.0000	3428972	0
perc_sem	0.3960	0.3010	0.0189	1.0000	3428972	0
NR_TRX	50.7037	64.4723	1.0000	2696.0000	3428972	0
cesta_rep	32.8463	27.9528	0.0054	9269.5915	3428972	0
idade	20213.8045	40900.7533	1.0000	735964.0000	3317224	111748
agregado	3.4874	6.3226	0.0000	99.0000	3328136	100836
lifetime	3046.9448	1422.5418	0.0000	4396.0000	3428867	105
vlr_resp	1134.1370	1470.7389	0.0000	76906.1874	3428972	0
primav	0.0077	0.0873	0.0000	1.0000	3428972	0
segm_baby_junior	0.2095	0.4069	0.0000	1.0000	3428972	0
segm_grocer	0.0015	0.0391	0.0000	1.0000	3428972	0
cart_uni	0.0574	0.2325	0.0000	1.0000	3428972	0
DL_SP	149.3722	241.5567	0.0000	61441.1271	3428972	0
vb_eco	127.0917	320.5799	0.0000	296321.0369	3428972	0
nr_ins	1.7349	1.6955	0.0000	8.0000	3428972	0
vb_nalim	180.2621	292.9119	0.0000	61210.1203	3428972	0
FLAG_CLOSER_PARTNER	0.1509	0.3579	0.0000	1.0000	3428972	0
OVERALL_STORES_DENSITY	0.8448	0.3374	0.1958	6.0809	3033219	395753
dl_tot	233.2893	332.5978	0.0000	62398.3943	3428972	0
perc_vb_eco	0.0975	0.1586	0.0000	1.0000	3428972	0
perc_vb_na	0.1554	0.1742	0.0000	1.0000	3428966	6
x	49.3874	64.4696	0.0000	2695.0000	3428972	0
t	271.9000	118.0418	0.0000	363.0000	3428972	0
T_cohort	311.9355	82.8930	0.0000	363.0000	3428972	0

Tabela 3 Principais medidas de centralidade e dispersão (cont.)

Variable	Lower Quartile	Median	Upper Quartile	Coeff of Variati
VLR	189.0726	615.8820	1548.9874	126.32
VB	232.5904	771.0691	1915.5985	125.12
VL	201.0165	668.5048	1667.1519	125.67
DESCONTOS	7.6978	44.2777	133.6534	145.46
t_recency	1.7330	9.1968	36.4354	178.11
perc_sem	0.1229	0.3420	0.6367	76.01
NR_TRX	10.6110	30.8927	66.9873	127.15
cesta_rep	16.1976	26.1253	41.8002	85.10
idade	13455.8461	17286.9407	22045.3121	202.34
agregado	2.0000	3.0000	4.0000	181.29
lifetime	1566.4486	3622.3476	4294.0197	46.68
vlr_resp	160.4032	609.6142	1580.8928	129.67
primav	0.0000	0.0000	0.0000	1136.88
segm_baby_junior	0.0000	0.0000	0.2831	194.26
segm_grocer	0.0000	0.0000	0.0000	2556.13
cart_uni	0.0000	0.0000	0.0001	405.40
DL_SP	8.3790	56.8795	193.6235	161.71
vb_eco	0.0049	32.5152	139.5146	252.24
nr_ins	0.0109	1.0007	3.0006	97.72
vb_nalim	18.8310	83.1440	228.5972	162.49
FLAG_CLOSER_PARTNER	0.0000	0.0000	0.0000	237.22
OVERALL_STORES_DENSITY	0.5556	0.8105	1.0688	39.93
dl_tot	26.3655	111.9751	312.7457	142.56
perc_vb_eco	0.0000	0.0327	0.1213	162.54
perc_vb_na	0.0408	0.1032	0.2055	112.09
x	9.2391	29.6626	65.5888	130.53
t	226.9991	334.1955	354.3608	43.41
T_cohort	306.0172	351.4359	360.3858	26.57

A base é composta por 3.428.792 clientes, mas algumas variáveis só têm informação para uma parte deles, que é o caso da idade, agregado, *lifetime*, overall_stores_density e perc_vb_na. No tratamento da base de dados, falar-se-á do procedimento aplicado a estes valores omissos.

A magnitude de valores, de variável para variável, varia bastante. Uma vez que os valores em estudo são relativos a um ano, todas as variáveis de venda apresentam valores muito elevados, com médias que rondam os 1200€. Nestas condições também está a variável idade, medida em dias[6], que apresenta média de 20.213,8045 dias, o que corresponde a 55 anos de idade. Por outro lado, temos variáveis em percentual e variáveis binárias, com valores entre 0 e 1.

[6] A idade foi medida em dias para ficar coerente com a restantes variáveis de tempo: t_recency, Lifetime, T_cohort e t.

Através destas duas tabelas também é possível concluir que a maioria das variáveis têm uma grande concentração de clientes nos valores mais baixos de cada variável. Esta conclusão é facilmente retirada quando se observam os valores dos quartis de cada variável na Tabela 3. Por exemplo, o 1º quartil da variável t_recency é de 1,73 dias e o 3º quartil de 36,45 dias, o que significa que 25% da população efetuou a sua compra a pouco menos de 2 dias antes do final do ano de 2015, e 75% a menos de aproximadamente 36 dias. O valor máximo registado na variável é de 363 dias, aproximadamente, por isso metade da população concentra-se abaixo de 10% do valor máximo da variável. Nas variáveis binárias, este facto ainda é mais evidente pois os valores dos três quartis são todos iguais a zero, ou muito próximos de zero. O histograma é a representação mais clara deste facto. Nas Figuras 1-4 estão representados todos os histogramas de todas as variáveis quantitativas, onde é possível observar a concentração da distribuição ao lado esquerdo. Uma distribuição com estas características classifica-se como uma distribuição com assimetria positiva.

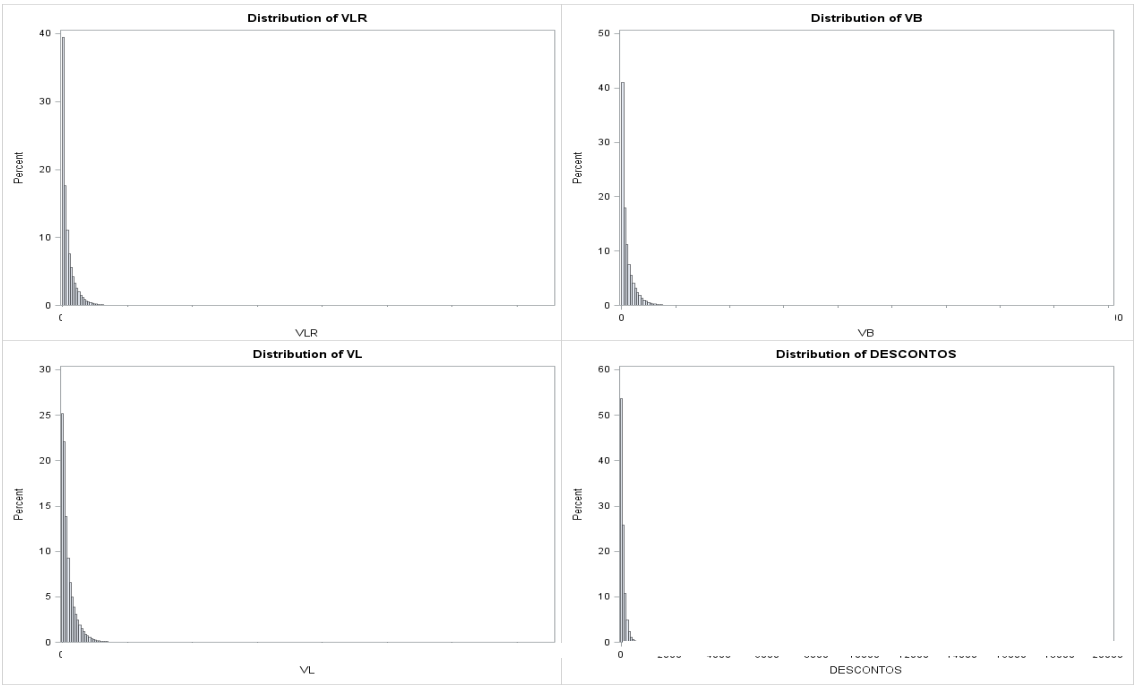


Figura 1 Histogramas das variáveis quantitativas VLR, VB, VL e DESCONTOS

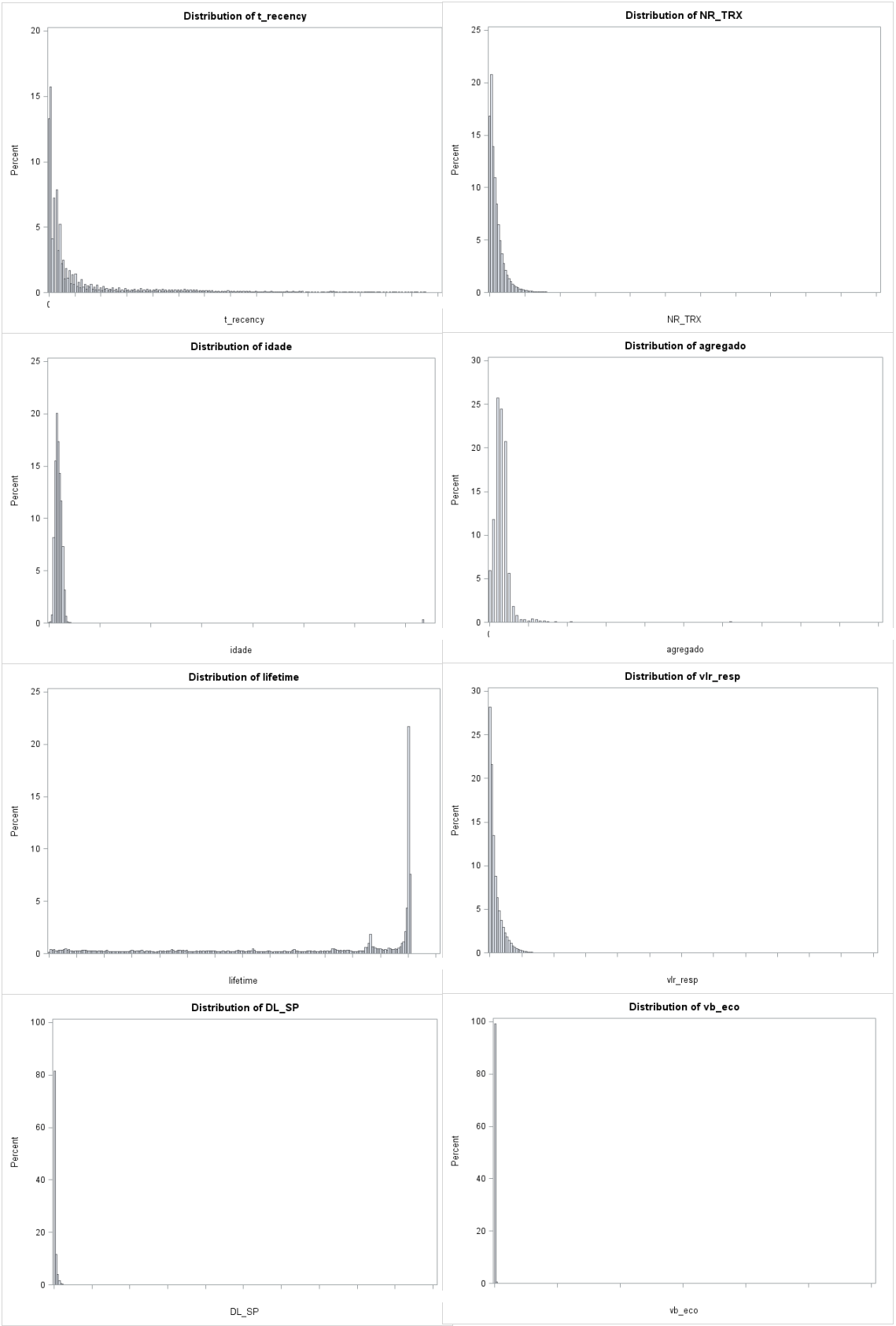


Figura 2 Histogramas das variáveis quantitativas t_recency, nr_trx, idade, agregado, *lifetime*, vlr_resp, dl_sp e vb_eco

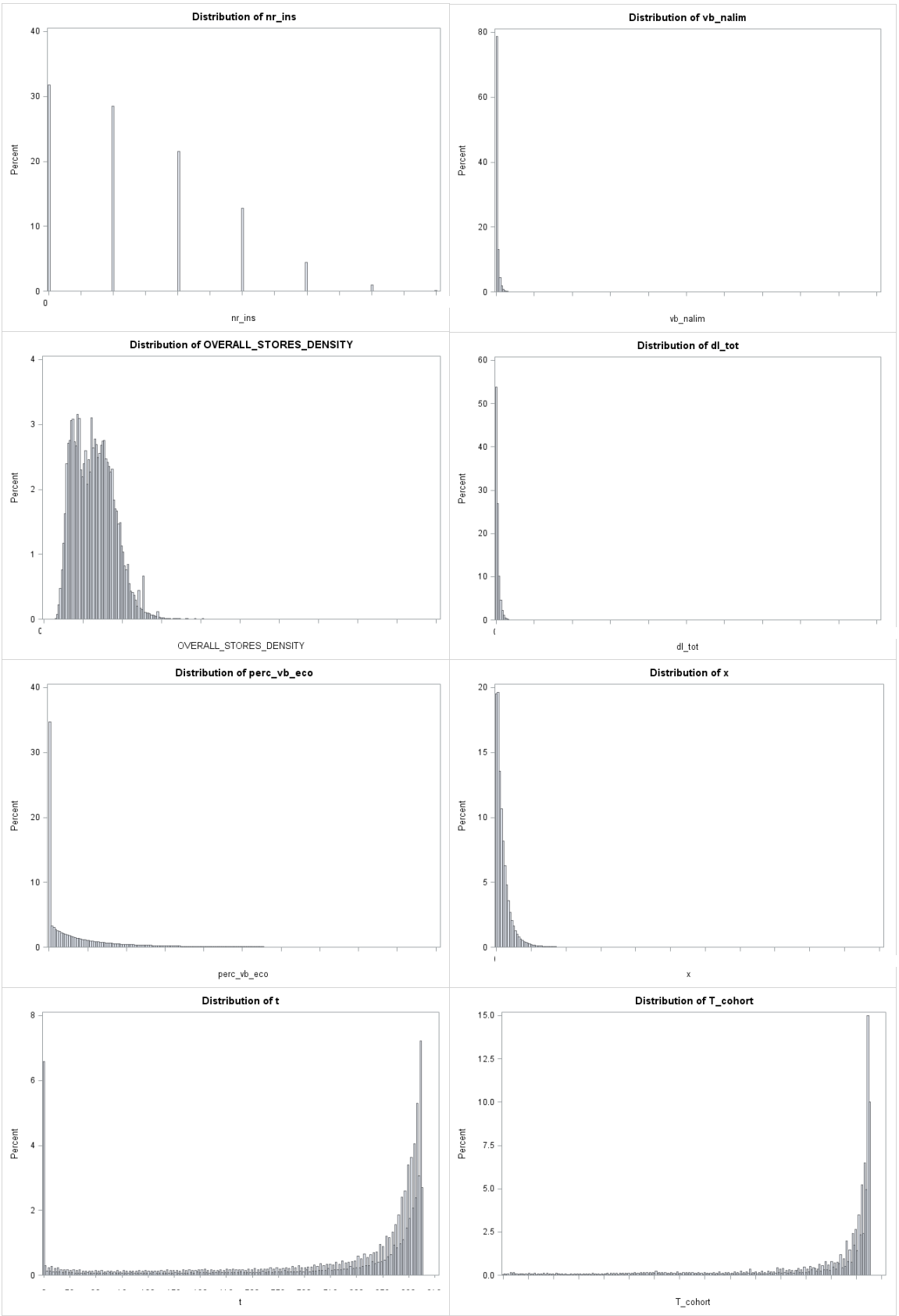


Figura 3 Histogramas das variáveis quantitativas nr_ins, vb_nalim, overall_stores_density, dl_tot, perc_vb_eco, x, t e T_cohort

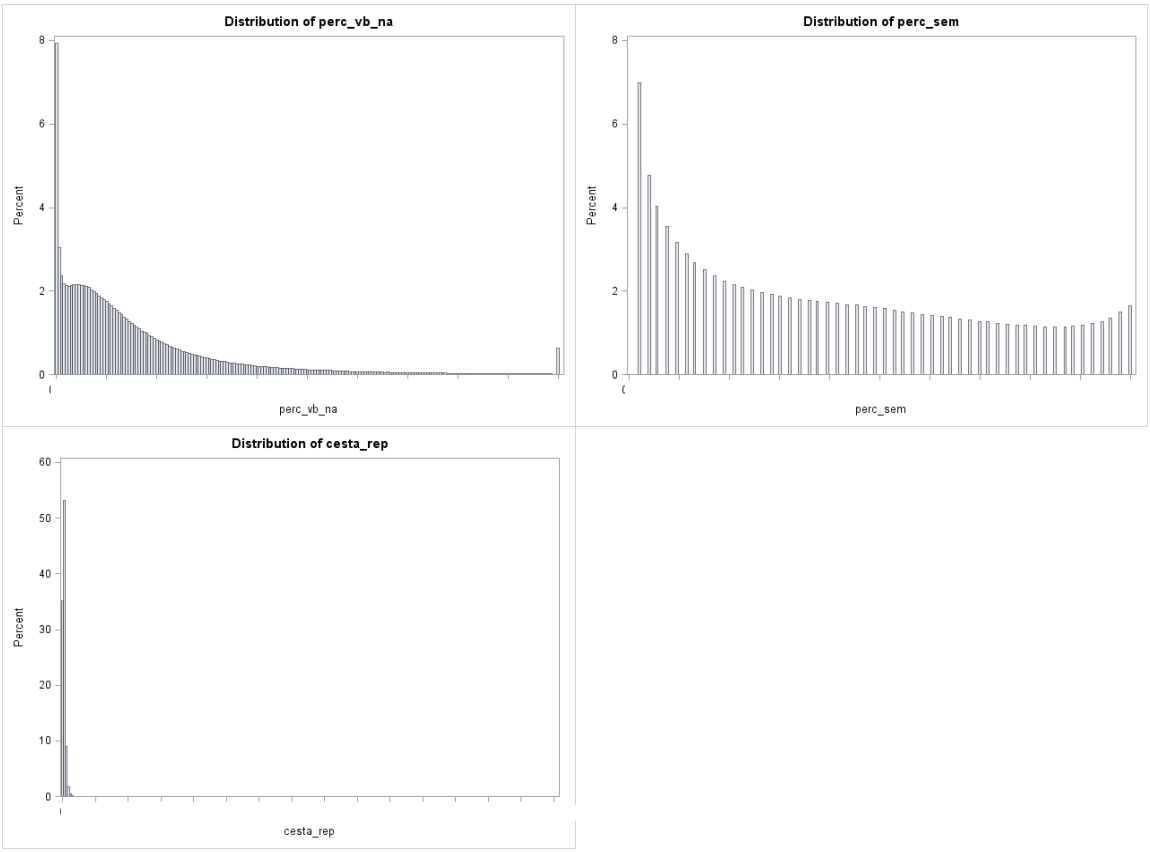


Figura 4 Histogramas das variáveis quantitativas perc_vb_na, perc_sem e cesta_rep

Apenas três das variáveis numéricas não têm o mesmo comportamento, apresentando uma distribuição com assimetria negativa, são elas o *lifetime*, o t e o T_cohort. Isto significa que uma grande percentagem de clientes tem uma conta antiga e que apresenta compras no início e no fim do ano de 2015, respetivamente.

As variáveis quantitativas binárias usadas para a previsão do CLV assinalam se o cliente pertence ou não a uma determinada condição. Estas condições correspondem a minorias que, quando observadas, indiciam um maior gasto do cliente, ou seja, um maior CLV.

Na Tabela 4 estão representadas as tabelas de frequência das cinco variáveis binárias que serão usadas como variáveis explicativas na regressão linear.

Tabela 4 Tabelas de frequência das variáveis explicativas binárias

primav	nº clientes	% clientes
0	3.402.646	99,2%
1	26.326	0,8%

segm_*baby_junior*	nº clientes	% clientes
0	2.710.712	79,1%
1	718.260	20,9%

segm_*grocer*	nº clientes	% clientes
0	3.423.732	99,8%
1	5.240	0,2%

cart_uni	nº clientes	% clientes
0	3.232.300	94,3%
1	196.672	5,7%

FLAG_CLOSER_PARTNER	nº clientes	% clientes
0	2.911.591	84,9%
1	517.381	15,1%

Em todas as variáveis é atribuído o valor 1 quando o cliente cumpre a condição. Assim tem-se que 0,8% dos clientes têm um desconto adicional nas suas compras, 20,9% têm crianças, 0,2% são comerciantes, 5,7% têm cartão de crédito com benefícios adicionais e 15,1% dos clientes têm uma loja da empresa mais próxima da sua residência do que uma loja da concorrência.

Após uma breve análise univariada das variáveis quantitativas, é importante olhar também para a única variável categórica utilizada no estudo: segm_valor. A distribuição dos clientes pelos segmentos encontra-se representada na Tabela 5.

Tabela 5 Tabela de frequência da variável categórica segm_valor

segm_valor	nº clientes	% clientes
Loyal	474.577	13,8%
Frequent	1.486.300	43,3%
Occasional	1.382.622	40,3%
Sem Valor	85.473	2,5%

O segmento moda desta variável é o segmento *Frequent,* com 43,3% dos clientes.

Análise Bivariada

Nos modelos preditivos, especialmente na Regressão Linear, é importante analisar a correlação entre as variáveis explicativas. A correlação pode ser calculada através de três coeficientes: Coeficiente de *Pearson*, de *Spearman* e *Phi*. Como se pretende calcular a correlação entre variáveis quantitativas o Coeficiente de *Pearson* é o mais adequado. Para variáveis ordinais deve utilizar-se o Coeficiente de *Spearman* e no caso das variáveis nominais deve utilizar-se o Coeficiente de *Phi* (Maroco, 2007). Recorrendo ao *software* SAS, calculou-se a correlação entre todas as variáveis quantitativas. O estudo de correlações será realizado entre variáveis de dois conjuntos: as vinte variáveis de entrada para a Regressão Linear e as quatro variáveis utilizadas no Modelo de Pareto/NBD. O resultado está representado na Tabela 6 e na Tabela 7, respetivamente.

Tabela 6 Matriz de Correlações entre as variáveis de entrada quantitativas da Regressão Linear

	vlr_resp	vlr	vb	vl	descontos	dl_sp	dl_tot	vb_eco	vb_nalim	perc_vb_eco	perc_vb_na	nr_ins	cesta_rep	t_recency	perc_sem	nr_trx	idade	agregado	lifetime	primav	segm_baby_junior	segm_grocer	cart_uni	flag_closer_partner	overall_stores_density
vlr_resp	**1,00**	**0,89**	**0,89**	**0,89**	**0,71**	**0,61**	**0,71**	**0,26**	**0,60**	**-0,12**	**-0,11**	**0,33**	**0,29**	**-0,33**	**0,66**	**0,60**	**0,01**	**0,02**	**0,22**	**0,11**	**0,23**	**-0,02**	**0,12**	**0,13**	**0,03**
vlr	**0,89**	1,00	1,00	1,00	0,79	0,69	0,79	0,28	0,67	-0,15	-0,12	0,36	0,33	-0,33	0,74	0,68	0,01	0,02	0,26	0,11	0,25	-0,03	0,11	0,13	0,04
vb	**0,89**	1,00	1,00	1,00	0,82	0,70	0,81	0,28	0,69	-0,15	-0,12	0,36	0,34	-0,34	0,74	0,67	0,01	0,02	0,26	0,11	0,25	-0,02	0,11	0,13	0,03
vl	**0,89**	1,00	1,00	1,00	0,81	0,70	0,81	0,28	0,68	-0,15	-0,12	0,36	0,33	-0,34	0,74	0,67	0,01	0,02	0,26	0,11	0,25	-0,02	0,11	0,13	0,04
descontos	**0,71**	0,79	0,82	0,81	1,00	0,63	0,83	0,24	0,68	-0,12	-0,04	0,32	0,31	-0,30	0,60	0,45	0,01	0,01	0,26	0,09	0,24	0,03	0,13	0,07	0,02
dl_sp	**0,61**	0,69	0,70	0,70	0,63	1,00	0,96	0,19	0,43	-0,11	-0,12	0,28	0,22	-0,26	0,56	0,49	-0,01	0,01	0,20	0,08	0,21	-0,01	-0,14	0,07	0,01
dl_tot	**0,71**	0,79	0,81	0,81	0,83	0,96	1,00	0,23	0,56	-0,13	-0,10	0,32	0,27	-0,30	0,63	0,52	0,00	0,01	0,24	0,09	0,24	0,00	-0,05	0,08	0,02
vb_eco	**0,26**	0,28	0,28	0,28	0,24	0,19	0,23	1,00	0,26	0,39	0,02	0,41	0,07	-0,14	0,25	0,21	-0,01	0,01	0,10	0,08	0,26	-0,01	0,05	0,02	-0,02
vb_nalim	**0,60**	0,67	0,69	0,68	0,68	0,43	0,56	0,26	1,00	-0,06	0,24	0,32	0,24	-0,25	0,49	0,44	-0,01	0,01	0,19	0,10	0,31	0,01	0,10	0,06	0,02
perc_vb_eco	**-0,12**	-0,15	-0,15	-0,15	-0,12	-0,11	-0,13	0,39	-0,06	1,00	0,13	0,38	-0,12	0,05	-0,15	-0,11	-0,02	0,00	-0,02	0,01	0,16	-0,02	-0,01	-0,06	-0,05
perc_vb_na	**-0,11**	-0,12	-0,12	-0,12	-0,04	-0,12	-0,10	0,02	0,24	0,13	1,00	0,01	-0,03	0,04	-0,17	-0,12	-0,02	-0,01	-0,02	0,00	0,13	0,12	0,01	-0,06	-0,01
nr_ins	**0,33**	0,36	0,36	0,36	0,32	0,28	0,32	0,41	0,32	0,38	0,01	1,00	0,02	-0,31	0,45	0,33	-0,03	0,01	0,20	0,09	0,40	-0,03	0,09	0,03	-0,03
cesta_rep	**0,29**	0,33	0,34	0,33	0,31	0,22	0,27	0,07	0,24	-0,12	-0,03	0,02	1,00	-0,03	-0,02	-0,10	0,02	0,01	0,04	-0,03	0,08	-0,01	0,01	-0,05	-0,09
t_recency	**-0,33**	-0,33	-0,34	-0,34	-0,30	-0,26	-0,30	-0,14	-0,25	0,05	0,04	-0,31	-0,03	1,00	-0,51	-0,33	0,00	-0,01	-0,17	-0,04	-0,19	0,04	-0,09	-0,08	-0,02
perc_sem	**0,66**	0,74	0,74	0,74	0,60	0,56	0,63	0,25	0,49	-0,15	-0,17	0,45	0,02	-0,51	1,00	0,80	0,00	0,01	0,31	0,12	0,25	-0,04	0,12	0,19	0,09
nr_trx	**0,60**	0,68	0,67	0,67	0,45	0,49	0,52	0,21	0,44	-0,11	-0,12	0,33	-0,10	-0,33	0,80	1,00	0,00	0,01	0,21	0,20	0,18	-0,02	0,11	0,20	0,11
idade	**0,01**	0,01	0,01	0,01	0,01	-0,01	0,00	-0,01	0,01	-0,02	-0,02	-0,03	0,02	0,00	0,00	0,00	1,00	-0,01	0,01	-0,01	-0,03	0,00	-0,01	0,00	0,02
agregado	**0,02**	0,02	0,02	0,02	0,01	0,01	0,01	0,01	0,01	0,00	-0,01	0,01	0,01	-0,01	0,01	0,01	-0,01	1,00	0,02	0,00	0,01	0,00	0,00	0,01	-0,02
lifetime	**0,22**	0,26	0,26	0,26	0,26	0,20	0,24	0,10	0,19	-0,02	-0,02	0,20	0,04	-0,17	0,31	0,21	0,01	0,02	1,00	0,03	0,11	-0,03	0,06	0,04	0,02
primav	**0,11**	0,11	0,11	0,11	0,09	0,08	0,09	0,08	0,10	0,01	0,00	0,09	-0,03	-0,04	0,12	0,20	-0,01	0,00	0,03	1,00	0,05	0,00	0,07	0,00	0,01
segm_baby_junior	**0,23**	0,25	0,25	0,25	0,24	0,21	0,24	0,26	0,31	0,16	0,13	0,40	0,08	-0,19	0,25	0,18	-0,03	0,01	0,11	0,05	1,00	0,00	0,04	0,02	-0,02
segm_grocer	**-0,02**	-0,03	-0,02	-0,02	0,03	-0,01	0,00	-0,01	0,01	-0,02	0,12	-0,03	-0,01	0,04	-0,04	-0,02	0,00	0,00	-0,03	0,00	0,00	1,00	0,00	-0,01	0,00
cart_uni	**0,12**	0,11	0,11	0,11	0,13	-0,14	-0,05	0,05	0,10	-0,01	0,01	0,09	0,01	-0,09	0,12	0,11	-0,01	0,00	0,06	0,07	0,04	0,00	1,00	0,00	0,02
flag_closer_partner	**0,13**	0,13	0,13	0,13	0,07	0,07	0,08	0,02	0,06	-0,06	-0,06	0,03	-0,05	-0,08	0,19	0,20	0,00	0,01	0,04	0,00	0,02	-0,01	0,00	1,00	0,01
overall_stores_density	**0,03**	0,04	0,03	0,04	0,02	0,01	0,02	-0,02	0,02	-0,05	-0,01	-0,03	-0,09	-0,02	0,09	0,11	0,02	-0,02	0,02	0,01	-0,02	0,00	0,02	0,01	1,00

Tabela 7 Matriz de Correlações entre as variáveis de entrada quantitativas do Modelo de Pareto/NBD

	vlr_resp	VLR	x	t	T_cohort
vlr_resp	**1**	**0,89**	**0,6**	**0,4**	**0,29**
VLR	**0,89**	1	0,68	0,45	0,36
x	**0,6**	0,68	1	0,45	0,36
t	**0,4**	0,45	0,45	1	0,8
T_cohort	**0,29**	0,36	0,36	0,8	1

Na primeira linha estão as correlações de todas as variáveis com a variável resposta. As matrizes encontram-se em tons de vermelho, amarelo e verde para ser mais fácil a identificação das correlações mais fortes. Duas variáveis dizem-se positivamente correlacionadas quando o coeficiente é próximo de 1 (vermelho) e negativamente correlacionadas se próximo de -1 (verde). Não há correlação linear quando o valor do coeficiente é aproximadamente zero (amarelo). Na diagonal da matriz estão as correlações das variáveis com elas próprias.

Inicialmente analisa-se a Matriz de Correlações entre as variáveis de entrada da Regressão Linear e posteriormente a Matriz de Correlações entre as variáveis de entrada do Modelo Pareto/NBD.

Na Tabela 6, concentrando o foco de análise nas correlações de todas as variáveis com a variável resposta conclui-se que a maioria das variáveis de valor de vendas influenciam positivamente o CLV do cliente, especificamente as vendas líquidas reportadas, as vendas líquidas, as vendas brutas, as vendas brutas no ecossistema e as vendas brutas no sector não alimentar. O número de visitas, a cesta média e percentagem de semanas com compra influenciam também positivamente a variável resposta, com uma correlação mais fraca relativamente à registada entre a variável resposta e as variáveis anteriormente referidas. Apenas a variável *recency* apresenta uma correlação negativa com a variável resposta. Este facto acontece porque quanto menor for o número de dias entre a última compra e o último dia de análise, ou seja, quanto mais recente for a última compra do cliente, maior é a probabilidade de ele gastar um maior valor no ano seguinte do que para um cliente cuja a sua última compra é muito antiga. Todas as restantes variáveis não apresentam uma correlação relevante com a variável resposta. Deve-se, então, estudar a hipótese de exclusão destas variáveis do modelo de previsão.

Ainda na Tabela 6 verifica-se uma elevada correlação entre três grupos de variáveis: o primeiro grupo composto pelas variáveis VLR, VB e VL, o segundo pelas variáveis Descontos, dl_sp e dl_tot e o terceiro pelas variáveis perc_sem e nr_trx. Os coeficientes de correlação indicam uma relação positiva com valores de 1, 0,96 e 0,8, respetivamente.

Um dos pré-requisitos da regressão linear é a não-correlação entre variáveis explicativas. Assim é necessário selecionar as variáveis mais importantes para a previsão que não apresentem correlação elevada. Deste modo, as variáveis eliminadas são, dentro do grupo de variáveis fortemente correlacionadas (correlação superior a 0,8), as que menos influenciam a variável resposta. No caso das variáveis VLR, VL e VB, a que possui uma correlação mais forte com a variável resposta é a VLR com 0,89309, ligeiramente superior à registada pelas VL (0,89305) e pelas VB (0,89277). Nas variáveis de desconto que os clientes têm no ato de compra, apenas se excluiu a variável Dl_tot, pois apresenta uma forte correlação com as variáveis Descontos (0,83) e Dl_sp (0,96). Por fim exclui-se a variável Nr_trx pois apresenta uma correlação de 0,8 com a variável perc_sem e é, das duas variáveis, a que tem menor correlação com a variável resposta. A Matriz de Correlações entre as vinte variáveis selecionadas está representada na Tabela 8.

Na Matriz de Correlações entre as quatro variáveis de entrada do Modelo Pareto/NBD, a variável que apresenta uma maior correlação com o CLV do cliente em 2016 é a VLR, que corresponde ao valor de vendas líquidas reportadas em 2015.

Na relação entre as variáveis explicativas apenas duas delas apresentam uma forte correlação positiva: a variável t e a variável T_cohort no valor de 0,8. Na seleção da variável a ser tomada em consideração na previsão do CLV usou-se o mesmo critério que foi utilizado na seleção de variáveis para a Regressão Linear, ou seja, das duas variáveis correlacionadas fica a que apresentar uma maior correlação com a variável resposta. Neste caso a variável t é a selecionada pois apresenta uma correlação de 0,40, superior à correlação de 0,29 com a variável T_cohort.

Tabela 8 Matriz de Correlações entre as variáveis quantitativas selecionadas

	vlr_resp	vlr	descontos	dl_sp	vb_eco	vb_nalim	perc_vb_eco	perc_vb_na	nr_ins	cesta_rep	t_recency	perc_sem	idade	agregado	lifetime	primav	segm_baby_junior	segm_grocer	cart_uni	flag_closer_partner	overall_stores_density
vlr_resp	**1,00**	**0,89**	**0,71**	**0,61**	**0,26**	**0,60**	**-0,12**	**-0,11**	**0,33**	**0,29**	**-0,33**	**0,66**	**0,01**	**0,02**	**0,22**	**0,11**	**0,23**	**-0,02**	**0,12**	**0,13**	**0,03**
vlr	**0,89**	1,00	0,79	0,69	0,28	0,67	-0,15	-0,12	0,36	0,33	-0,33	0,74	0,01	0,02	0,26	0,11	0,25	-0,03	0,11	0,13	0,04
descontos	**0,71**	0,79	1,00	0,63	0,24	0,68	-0,12	-0,04	0,32	0,31	-0,30	0,60	0,01	0,01	0,26	0,09	0,24	0,03	0,13	0,07	0,02
dl_sp	**0,61**	0,69	0,63	1,00	0,19	0,43	-0,11	-0,12	0,28	0,22	-0,26	0,56	-0,01	0,01	0,20	0,08	0,21	-0,01	-0,14	0,07	0,01
vb_eco	**0,26**	0,28	0,24	0,19	1,00	0,26	0,39	0,02	0,41	0,07	-0,14	0,25	-0,01	0,01	0,10	0,08	0,26	-0,01	0,05	0,02	-0,02
vb_nalim	**0,60**	0,67	0,68	0,43	0,26	1,00	-0,06	0,24	0,32	0,24	-0,25	0,49	-0,01	0,01	0,19	0,10	0,31	0,01	0,10	0,06	0,02
perc_vb_eco	**-0,12**	-0,15	-0,12	-0,11	0,39	-0,06	1,00	0,13	0,38	-0,12	0,05	-0,15	-0,02	0,00	-0,02	0,01	0,16	-0,02	-0,01	-0,06	-0,05
perc_vb_na	**-0,11**	-0,12	-0,04	-0,12	0,02	0,24	0,13	1,00	0,01	-0,03	0,04	-0,17	-0,02	-0,01	-0,02	0,00	0,13	0,12	0,01	-0,06	-0,01
nr_ins	**0,33**	0,36	0,32	0,28	0,41	0,32	0,38	0,01	1,00	0,02	-0,31	0,45	-0,03	0,01	0,20	0,09	0,40	-0,03	0,09	0,03	-0,03
cesta_rep	**0,29**	0,33	0,31	0,22	0,07	0,24	-0,12	-0,03	0,02	1,00	-0,03	-0,02	0,02	0,01	0,04	-0,03	0,08	-0,01	0,01	-0,05	-0,09
t_recency	**-0,33**	-0,33	-0,30	-0,26	-0,14	-0,25	0,05	0,04	-0,31	-0,03	1,00	-0,51	0,00	-0,01	-0,17	-0,04	-0,19	0,04	-0,09	-0,08	-0,02
perc_sem	**0,66**	0,74	0,60	0,56	0,25	0,49	-0,15	-0,17	0,45	-0,02	-0,51	1,00	0,00	0,01	0,31	0,12	0,25	-0,04	0,12	0,19	0,09
idade	**0,01**	0,01	0,01	-0,01	-0,01	-0,01	-0,02	-0,02	-0,03	0,02	0,00	0,00	1,00	-0,01	0,01	-0,01	-0,03	0,00	-0,01	0,00	0,02
agregado	**0,02**	0,02	0,01	0,01	0,01	0,01	0,00	-0,01	0,01	0,01	-0,01	0,01	-0,01	1,00	0,02	0,00	0,01	0,00	0,00	0,01	-0,02
lifetime	**0,22**	0,26	0,26	0,20	0,10	0,19	-0,02	-0,02	0,20	0,04	-0,17	0,31	0,01	0,02	1,00	0,03	0,11	-0,03	0,06	0,04	0,02
primav	**0,11**	0,11	0,09	0,08	0,08	0,10	0,01	0,00	0,09	-0,03	-0,04	0,12	-0,01	0,00	0,03	1,00	0,05	0,00	0,07	0,00	0,01
segm_baby_junior	**0,23**	0,25	0,24	0,21	0,26	0,31	0,16	0,13	0,40	0,08	-0,19	0,25	-0,03	0,01	0,11	0,05	1,00	0,00	0,04	0,02	-0,02
segm_grocer	**-0,02**	-0,03	0,03	-0,01	-0,01	0,01	-0,02	0,12	-0,03	-0,01	0,04	-0,04	0,00	0,00	-0,03	0,00	0,00	1,00	0,00	-0,01	0,00
cart_uni	**0,12**	0,11	0,13	-0,14	0,05	0,10	-0,01	0,01	0,09	0,01	-0,09	0,12	-0,01	0,00	0,06	0,07	0,04	0,00	1,00	0,00	0,02
flag_closer_partner	**0,13**	0,13	0,07	0,07	0,02	0,06	-0,06	-0,06	0,03	-0,05	-0,08	0,19	0,00	0,01	0,04	0,00	0,02	-0,01	0,00	1,00	0,01
overall_stores_density	**0,03**	0,04	0,02	0,01	-0,02	0,02	-0,05	-0,01	-0,03	-0,09	-0,02	0,09	0,02	-0,02	0,02	0,01	-0,02	0,00	0,02	0,01	1,00

3.1.2 Pré-Processamento da Base de Dados

Algumas das variáveis da base de dados apresentam valores incoerentes tais como a variável agregado familiar com 99 membros ou a idade com o valor de 735.964 dias que corresponde a 2014,26 anos, facto impossível na vida de um ser humano. De forma a evitar que estas observações causem algum tipo de ruído nos resultados dos modelos, estas serão eliminadas. No caso do agregado familiar, uma vez que em Portugal, a média do número de membros de agregados familiares com mais do que 6 membros é aproximadamente 19,4 (INE, 2017) e assumindo que esta variável tem uma distribuição simétrica, foi considerado aceitável um número máximo de 19,4+6≈32 membros. Na idade, dado que o ser humano que registou a duração mais longa de vida atingiu os 122 anos (44.560 dias) (Eide, 2016), esse foi o corte efetuado na variável, ou seja, todos os clientes que apresentavam mais do que 44.560 dias de idade foram eliminados.

Após a exclusão dos dados incoerentes iniciou-se o estudo dos *outliers*. Existem 2 tipos de *outliers*: os moderados e os severos. As observações são consideradas *outliers* moderados se estiverem entre $Q_1 - 1{,}5D$ e $Q_1 - 3D$ ou entre $Q_3 + 1{,}5D$ e $Q_3 + 3D$, com $D = Q_3 - Q_1$ como a Amplitude Inter-quartis, Q_1valor do primeiro quartil e Q_3 valor do terceiro quartil. Se forem menores do que $Q_1 - 3D$ ou maiores do que $Q_3 + 3D$ são

considerados *outliers* severos. Na identificação dos *outliers* presentes na base dados optou-se por apenas identificar os *outliers* severos de modo a evitar um corte drástico de observações e uma possível perda de informação.

Nas Figuras 5-7 estão representados os *box-plots* das vinte variáveis numéricas da base de dados em estudo. É possível observar que a maioria delas apresentam *outliers* exceto a perc_sem, lifetime e nr_ins. Também é possível concluir que todos os *outliers* observados são superiores. Assim o estudo destes *outliers* será concentrado apenas nas observações que apresentavam valores superiores a $Q_3 + 3D$ e em variáveis quantitativas contínuas ou discretas.

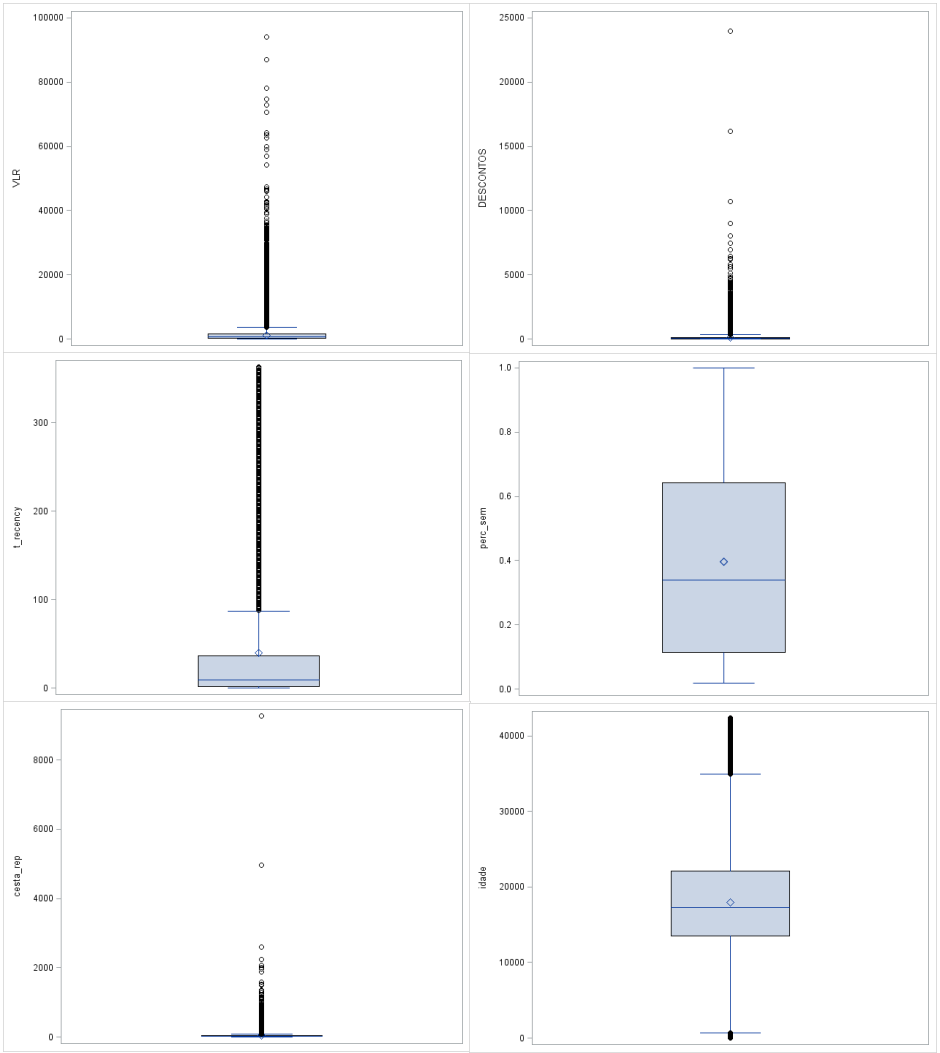


Figura 5 *Box-plots* das variáveis quantitativas vlr, descontos, t_recency, perc_sem, cesta_rep e idade

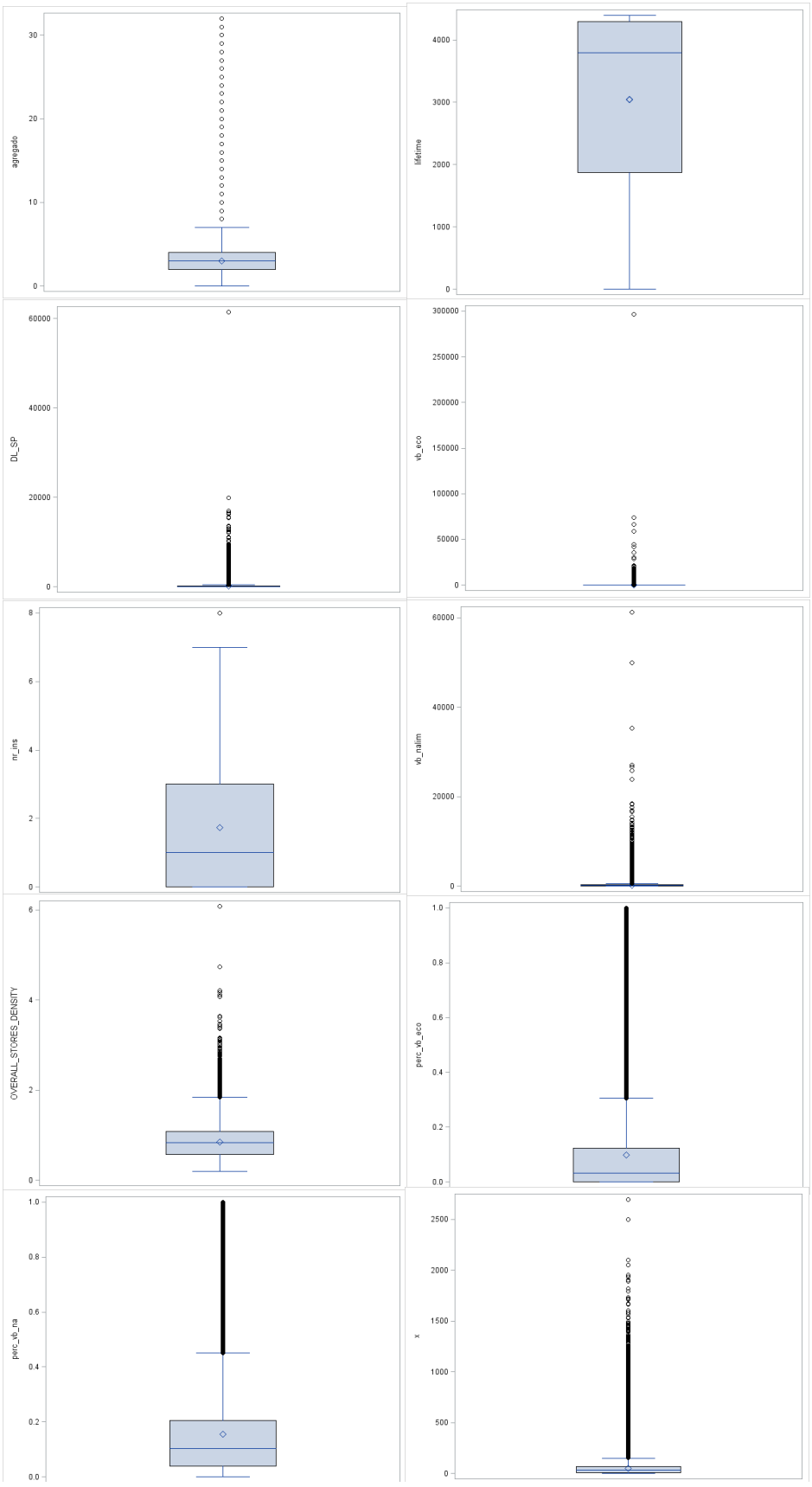


Figura 6 *Box-plots* das variáveis quantitativas agregado, *lifetime*, dl_sp, vb_eco, nr_ins, vb_nalim, overall_stores_density, perc_vb_eco, perc_vb_na e x

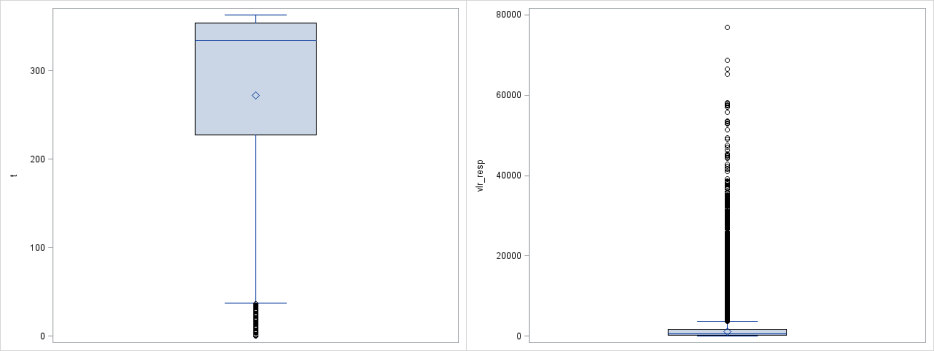


Figura 7 *Box-plots* das variáveis quantitativas t e vlr_resp

Avaliou-se a existência de observações *outliers* severos nas variáveis vlr, descontos, cesta_rep, dl_sp, vb_eco, vb_nalim, overall_stores_density, x e t, incluindo a variável resposta vlr_resp. O resultado apresentado indica, por cada cliente, em quantas variáveis é considerado *outlier*.

Tabela 9 Nº de Clientes por nº de variáveis onde são considerados *outliers* (n)

n	Nº clientes	vlr	descontos	cesta_rep	dl_sp	vb_eco	vb_nalim	overall_stores_densi	vlr_resp	x	t	% clientes
0	2.980.781	0	0	0	0	0	0	0	0	0	0	87,9
1	283.096	2.139	10.322	38.013	38.064	117.112	35.796	967	5.398	35.285	0	8,4
2	64.154	7.499	19.358	4.201	22.830	23.761	27.439	99	8.182	14.939	0	1,9
3	29.066	12.581	13.877	1.996	13.876	10.295	16.073	12	11.075	7.413	0	0,9
4	17.455	13.773	10.355	1.499	9.704	6.275	10.471	12	12.244	5.487	0	0,5
5	10.287	9.858	7.819	1.554	7.045	4.434	7.922	6	9.188	3.609	0	0,3
6	4.196	4.178	3.740	937	3.529	2.703	3.942	2	4.085	2.060	0	0,1
7	789	789	788	234	789	784	788	2	789	560	0	0,0
8	4	4	4	4	4	4	4	0	4	4	0	0,0

Observando a Tabela 9, tem-se que 87,9% dos clientes não são considerados *outliers*, e quando são, a maioria dos clientes é considerado *outlier* apenas numa variável. Isolaram-se estes últimos clientes, para perceber o porquê de serem considerados *outliers*, se se trata de um erro de medição ou se representa um comportamento natural que se deve investigar. Assim cruzou-se o número de variáveis em que um cliente é considerado *outlier* (n) com os segmentos valor para perceber de que tipo de clientes se trata. A Tabela 10 apresenta o resultado deste cruzamento que permite concluir que os clientes considerados *outliers* são os clientes mais fieis à empresa. Quanto maior for o número de

variáveis onde o cliente é considerado *outlier* maior é o percentual de clientes no segmento *Loyal* da Segmentação Valor.

Tabela 10 Percentagem de clientes por nº de variáveis onde é classificado como *outlier* e segmento valor

% Clientes	Segmentos Valor				
Nº variáveis onde é considerado *outlier*	*Loyal*	*Frequent*	*Occasional*	Sem Valor	Grand Total
0	8,5%	44,0%	45,1%	2,5%	100,0%
1	41,5%	47,4%	8,5%	2,6%	100,0%
2	71,2%	25,6%	0,3%	2,8%	100,0%
3	80,8%	16,1%	0,1%	3,0%	100,0%
4	86,9%	9,8%	0,0%	3,3%	100,0%
5	88,4%	9,4%	0,0%	2,1%	100,0%
6	89,9%	8,9%	0,0%	1,1%	100,0%
7	92,9%	6,7%	0,0%	0,4%	100,0%
8	100,0%	0,0%	0,0%	0,0%	100,0%

Nesta dissertação pretende-se prever o valor do cliente para todas as classes de clientes, desde os que fazem compras diariamente aos clientes que compram ocasionalmente. Por isso nenhum *outlier* será excluído da base de dados, nesta fase, de modo a evitar a perda de informação dos melhores clientes da empresa.

Outro dos problemas muito frequentes com que os analistas se deparam quando trabalham com base de dados são os valores omissos ou *missings*. Este é um tema muito controverso na literatura pois existem imensos métodos para contornar este problema, uns mais complexos do que outros. Neste contexto, dado que as variáveis onde este evento se verifica não apresentam uma elevada variância (idade, agregado, *lifetime*, overall_stores_density e perc_vb_na) e que o percentual de *missings* é muito baixo (Tabela 11), optou-se pela substituição destes valores omissos por uma constante.

As variáveis em questão não aparentam seguir distribuição normal, assim, a constante que substituirá os valores omissos será a respetiva mediana, representando melhor a tendência central das mesmas (Veroneze, 2011).

Nas Figuras 8 e 9 pode-se ver que o impacto da imputação foi reduzido na distribuição das variáveis, mantendo-se os valores das medidas de centralidade e de dispersão próximos dos registados anteriormente.

Tabela 11 Número e percentual de *missings* por variável

Variável	N	N Missings	% Missings
vlr	3.389.828	0	0,00%
descontos	3.389.828	0	0,00%
t_recency	3.389.828	0	0,00%
perc_sem	3.389.828	0	0,00%
cesta_rep	3.389.828	0	0,00%
idade	3.278.574	111.254	3,28%
agregado	3.289.016	100.812	2,97%
lifetime	3.389.723	105	0,00%
vlr_resp	3.389.828	0	0,00%
primav	3.389.828	0	0,00%
segm_baby_junior	3.389.828	0	0,00%
segm_grocer	3.389.828	0	0,00%
cart_uni	3.389.828	0	0,00%
dl_sp	3.389.828	0	0,00%
vb_eco	3.389.828	0	0,00%
nr_ins	3.389.828	0	0,00%
vb_nalim	3.389.828	0	0,00%
flag_closer_partner	3.389.828	0	0,00%
overall_stores_density	2.997.955	391.873	11,56%
perc_vb_eco	3.389.828	0	0,00%
perc_vb_na	3389822	6	0,00%
x	3389828	0	0,00%
t	3389828	0	0,00%
t_cohort	3389828	0	0,00%

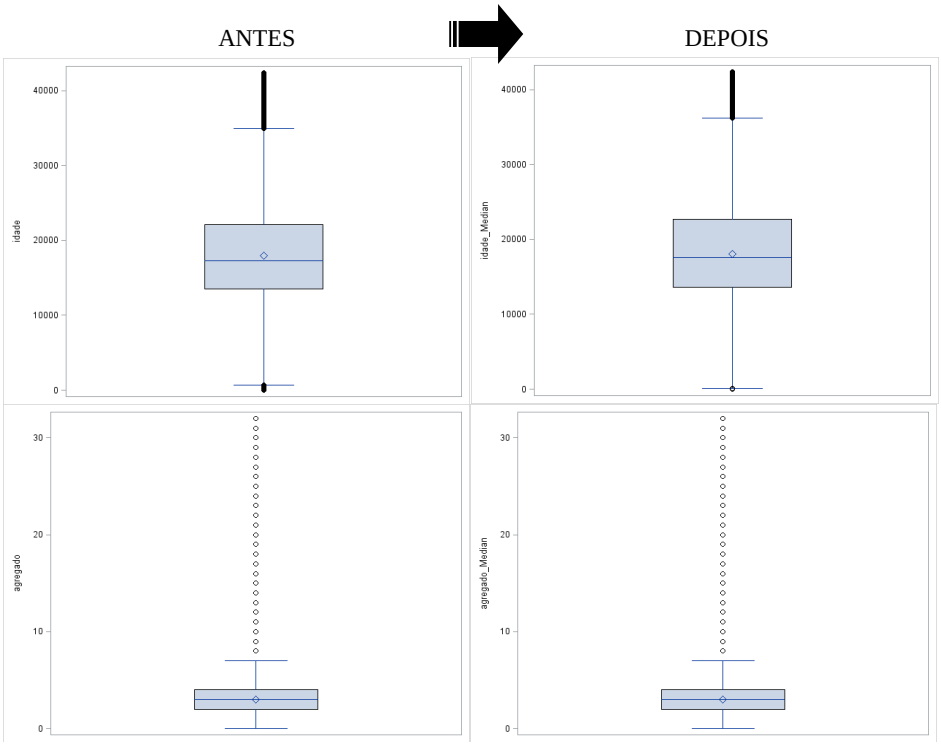


Figura 8 *Box-plots* das variáveis idade de agregado, antes e depois da imputação

ANTES ➡ DEPOIS

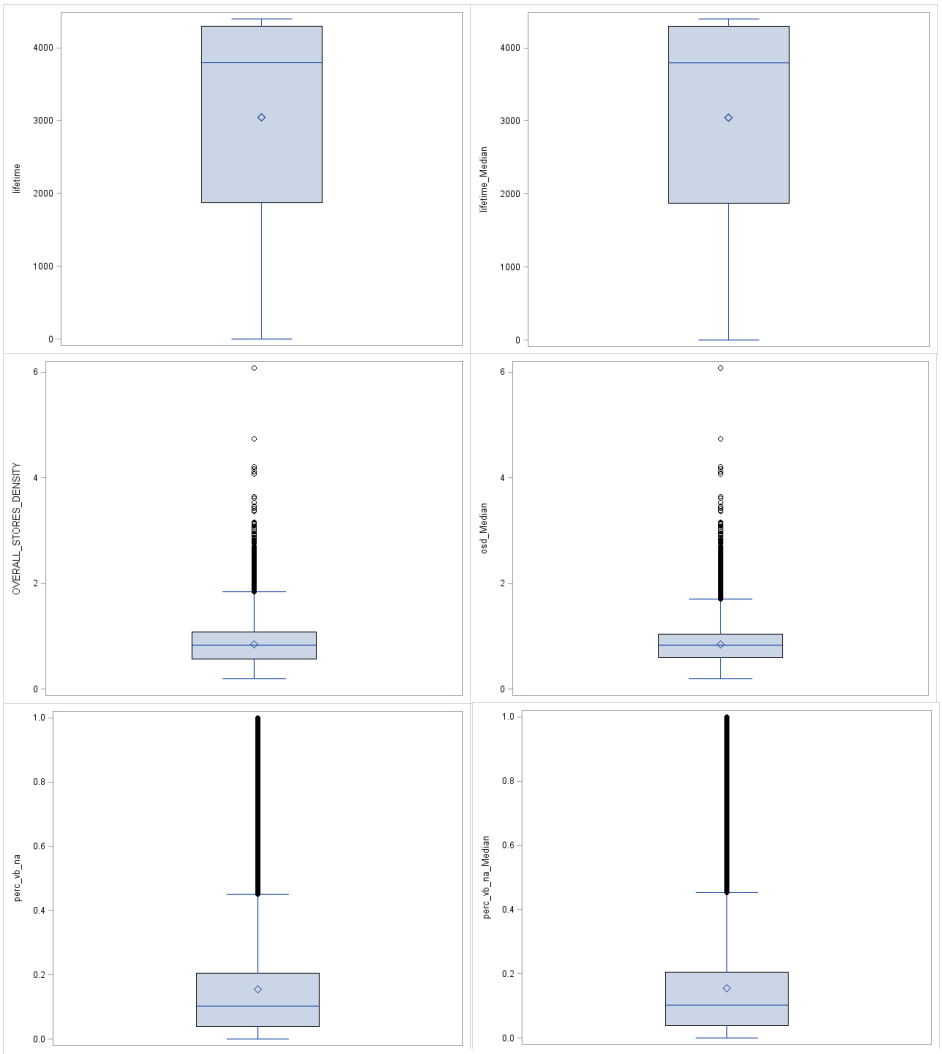


Figura 9 *Box-plots* das variáveis *lifetime*, *overall_stores_density* e perc_vb_na, antes e depois da imputação

4 Previsão do CLV

A previsão do CLV será calculada recorrendo a três modelos: Regressão Linear com variáveis de negócio, Modelo de Pareto/NBD e Regressão Linear com as variáveis de entrada do Modelo de Pareto/NBD.

Neste capítulo descreve-se todas as etapas do desenvolvimento dos três modelos, validação de pressupostos e respetivos resultados.

Devido à base de dados em estudo tem uma dimensão na ordem dos milhões, todos testes de hipóteses não podem ser usados para retirar conclusões. Tal como apresentado por Lin (2013) o gráfico Coefficient/p-value/sample-size (CPS) demonstra claramente o impacto da dimensão da amostra no p-value está representado na Figura 10.

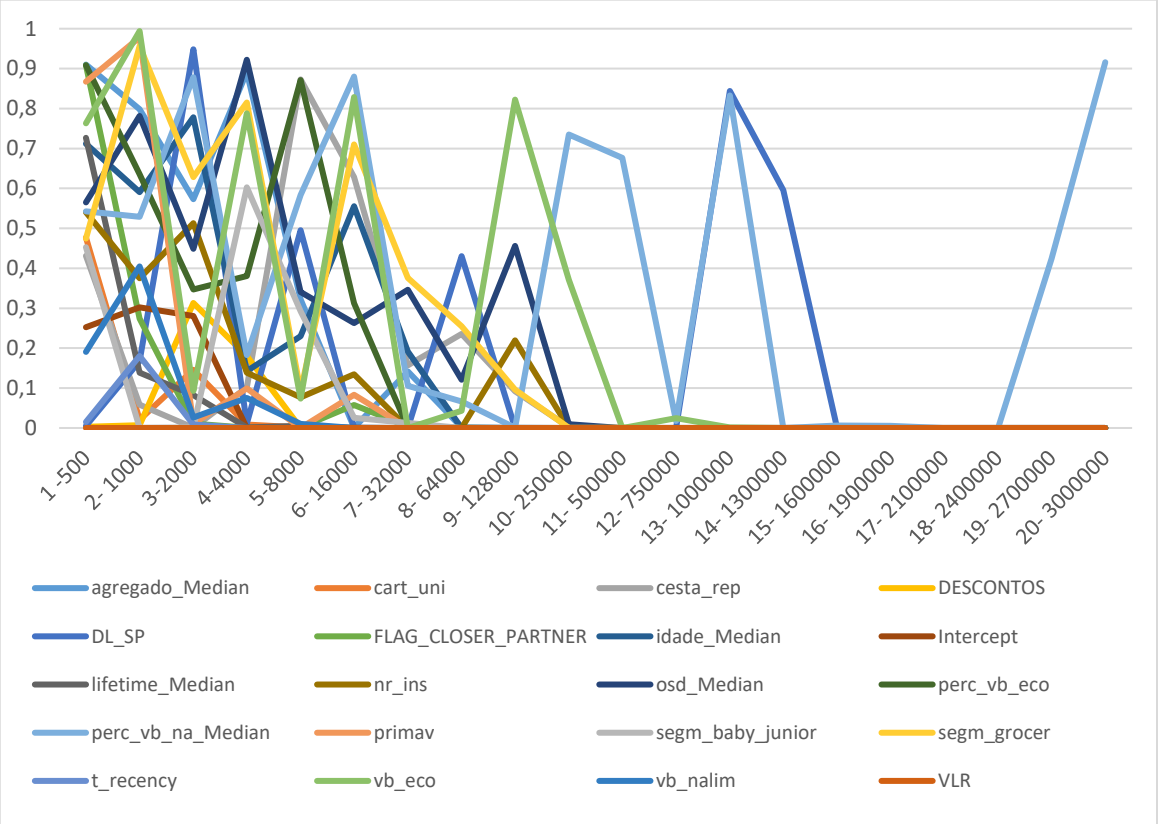


Figura 10 CPS *Chart*

Através do gráfico é possível verificar que quanto maior for a dimensão do conjunto de dados analisado, mais próximo de 0 fica o p-value do teste da significância dos coeficientes estimados.

4.1 Regressão Linear Múltipla

A previsão do CLV, através da Regressão Linear Múltipla, será desenvolvida com recurso ao *software* SAS. As variáveis explicativas que serão utilizadas neste modelo já foram referenciadas no Capítulo 3. No desenvolvimento deste modelo deve-se ter em conta o cumprimento dos pressupostos para que os resultados, e consequentes interpretações, sejam corretas. O estudo da correlação entre variáveis independentes efetuado no capítulo anterior permite a exclusão de algumas variáveis, à partida, que poderiam criar ruído no modelo. Mas estudos adicionais serão necessários para melhorar o modelo e obter melhores previsões.

Primeiramente, a previsão do CLV será modelada tendo por base o conjunto completo de observações, de forma a perceber como o modelo se comporta, qual a percentagem de variância explicada, o erro preditivo e validação dos pressupostos. Esta abordagem levanta dois problemas: os testes de hipóteses inerentes ao modelo não serão sensíveis a eventuais problemas existentes devido ao elevado número de observações em análise e o processamento dos resultados será demorado e pesado para o sistema.

De seguida, o modelo será aplicado a amostras estratificadas extraídas da população para a obtenção de previsões que serão comparadas posteriormente com as previsões obtidas pelo Modelo de Pareto/NBD.

Por fim, também será desenvolvido um modelo de Regressão Linear com as variáveis de entrada utilizadas no Modelo de Pareto/NBD, para uma melhor comparação de erros.

4.1.1 Regressão Linear para a Conjunto Total

Numa primeira fase utilizaram-se todas as variáveis, incluindo as que foram anteriormente identificadas como correlacionadas para sustentar a exclusão de algumas variáveis correlacionadas.

Na Figura 11 é apresentado o *output* da Regressão Linear do *software* SAS onde é indicado o número de observações em estudo, a ANOVA e algumas medidas de desempenho do modelo, tais como o erro quadrático médio (RMSE), a média da variável resposta, o coeficiente de variação dos erros, o coeficiente de determinação (R^2) e o coeficiente de determinação ajustado (R_a^2). A ANOVA testa se os coeficientes estimados da regressão linear são todos nulos ou se existe pelo menos um que não respeite essa condição. Tratando-se de um teste de hipóteses não é possível retirar conclusões com a base de dados em estudo. O R^2 e o R_a^2 apresentam o mesmo valor, de 0.8013, que significa que o modelo explica 80,13% da variância do CLV. O RMSE é de 496,13 o que significa que, em média, a previsão do modelo se distancia do valor real em 496,13€. Todos estes resultados indicam que o modelo tem qualidade.

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: vlr_resp

Number of Observations Read	3389828
Number of Observations Used	3389828

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	23	3.363981E12	1.4626E11	594192	<.0001
Error	3.39E6	8.343991E11	246150		
Corrected Total	3.39E6	4.19838E12			

Root MSE	496.13471	R-Square	0.8013
Dependent Mean	858.71384	Adj R-Sq	0.8013
Coeff Var	57.77649		

Figura 11 Regressão Linear com todas as variáveis - ANOVA e principais medidas de desempenho do modelo

Os resultados da estimação dos coeficientes da regressão estão apresentados na Figura 12, assim como o teste da significância de cada coeficiente, a Tolerância e a *Variance Inflaction Factor* (VIF), que são medidas utilizadas na análise da multicolinearidade.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	Intercept	1	176.81203	1.52230	116.15	<.0001	0	.	0
VLR	VLR	1	-6.62824	0.20307	-32.64	<.0001	-6.40567	0.00000152	656888
VB	VB	1	0.66274	0.01391	47.64	<.0001	0.78252	0.00021727	4602.47368
VL	VL	1	6.79020	0.19114	35.53	<.0001	7.01586	0.00000150	665227
DESCONTOS	DESCONTOS	1	-6.00848	0.17593	-34.15	<.0001	-0.57358	0.00020786	4811.02791
t_recency		1	-0.61112	0.00448	-136.27	<.0001	-0.03919	0.70903	1.41038
perc_sem		1	-91.90916	2.03215	-45.23	<.0001	-0.02486	0.19404	5.15347
NR_TRX	NR_TRX	1	-0.12190	0.01046	-11.65	<.0001	-0.00535	0.27791	3.59830
cesta_rep		1	-0.80238	0.01588	-50.54	<.0001	-0.01526	0.64314	1.55487
primav		1	199.32298	3.17177	62.84	<.0001	0.01567	0.94288	1.06058
segm_baby_junior		1	25.51656	0.75572	33.76	<.0001	0.00933	0.76746	1.30300
segm_grocer		1	-48.19393	7.01743	-6.87	<.0001	-0.00169	0.96570	1.03552
cart_uni		1	96.93460	1.24857	77.64	<.0001	0.02029	0.85853	1.16478
vb_eco		1	0.01896	0.00132	14.35	<.0001	0.00415	0.70153	1.42545
nr_ins		1	-2.10874	0.30636	-6.88	<.0001	-0.00233	0.51345	1.94762
vb_nalim		1	-0.13919	0.00213	-65.28	<.0001	-0.02776	0.32420	3.08450
FLAG_CLOSER_PARTNER		1	41.63058	0.77682	53.59	<.0001	0.01338	0.94082	1.06291
dl_tot		1	-0.06551	0.00228	-28.75	<.0001	-0.01484	0.22021	4.54119
perc_vb_eco		1	25.25444	2.17781	11.60	<.0001	0.00360	0.60873	1.64275
idade_Median		1	-0.00133	0.00005081	-26.16	<.0001	-0.00685	0.85565	1.16871
agregado_Median		1	2.27940	0.12510	18.22	<.0001	0.00443	0.99054	1.00955
lifetime_Median		1	-0.01157	0.00027416	-42.19	<.0001	-0.01119	0.83306	1.20039
osd_Median		1	-5.26340	1.13670	-4.63	<.0001	-0.00114	0.97177	1.02905
perc_vb_na_Median		1	-13.22135	1.78562	-7.40	<.0001	-0.00207	0.74997	1.33338

Figura 12 Regressão Linear com todas as variáveis - Parâmetros estimados, teste de significância dos coeficientes, coeficientes estimados estandardizados, Tolerância e VIF

Valores de VIF superiores a 5 e valores de Tolerância muito próximos de zero indicam a existência de multicolinearidade entre as variáveis independentes. Na Figura 12 observa-se que existem variáveis com valores de VIF na ordem dos milhares, especificamente nas variáveis VLR, VL, VB, descontos e perc_sem. Tratam-se das mesmas variáveis que apresentavam correlações elevadas na matriz de correlações analisadas no Capítulo 3.

O próximo passo é excluir as variáveis que foram indicadas com a análise da matriz de correlações, ou seja, VL, VB, nr_trx e dl_tot, e analisar também as restantes variáveis, a relação entre elas e a sua significância para a previsão do CLV.

Na seleção das variáveis a serem consideradas no modelo deve-se garantir que não existem variáveis que podem ser previstas como combinações lineares de outras variáveis. Para além da *Variance Inflaction Factor* (VIF) e da Tolerância, existem outras medidas que são indicadores da existência ou ausência de multicolinearidade entre as variáveis explicativas, que são também valores resultantes da análise feita no *software*

SAS, os *Eigenvalues* e o *Condition Index* (Petrini, Zulini, & Dias, 1999). As variáveis a incluir no modelo também devem ser significativas para a previsão do CLV. Neste âmbito, o *stepwise* foi o método de seleção aplicado ao modelo de Regressão Linear, com um nível de significância de entrada do modelo de 0.1 e um nível de significância de permanência no modelo de 0.01. Estes níveis são baixos relativamente aos que são normalmente praticados porque no tratamento de amostras de elevada dimensão devemos considerar um nível de significância inferior (Lin et al., 2013).

O *output* do SAS para a Regressão Linear com as variáveis independentes (sem pares de variáveis com elevada correlação), com seleção *stepwise* está apresentado na Tabela 12.

Tabela 12 Regressão Linear com todas as variáveis independentes – Tabela resumo da seleção *stepwise*

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	VLR		VLR	1	0.7978	0.7978	56954.6	1.337E7	<.0001
2	t_recency			2	0.0011	0.7989	37445.0	19298.5	<.0001
3	cart_uni			3	0.0006	0.7995	27543.8	9823.33	<.0001
4	DESCONTOS		DESCONTOS	4	0.0003	0.7998	22098.8	5411.81	<.0001
5	vb_nalim			5	0.0002	0.8000	18028.1	4051.15	<.0001
6	primav			6	0.0002	0.8003	13967.2	4046.25	<.0001
7	lifetime_Median			7	0.0002	0.8005	10871.6	3087.63	<.0001
8	FLAG_CLOSER_PARTNER			8	0.0001	0.8006	8760.07	2108.13	<.0001
9	perc_sem			9	0.0001	0.8007	6635.82	2122.11	<.0001
10	cesta_rep			10	0.0001	0.8008	4610.39	2024.68	<.0001
11	segm_baby_junior			11	0.0001	0.8010	2559.76	2051.09	<.0001
12	idade_Median			12	0.0000	0.8010	1732.50	828.84	<.0001
13	DL_SP			13	0.0000	0.8010	1134.09	600.21	<.0001
14	vb_eco			14	0.0000	0.8011	656.168	479.83	<.0001
15	agregado_Median			15	0.0000	0.8011	323.484	334.65	<.0001
16	segm_grocer			16	0.0000	0.8011	179.581	145.90	<.0001
17	perc_vb_eco			17	0.0000	0.8011	106.387	75.19	<.0001
18	perc_vb_na_Median			18	0.0000	0.8011	63.7925	44.59	<.0001
19	osd_Median			19	0.0000	0.8011	43.1013	22.69	<.0001
20	nr_ins			20	0.0000	0.8011	21.0000	24.10	<.0001

Analisando os resultados da regressão linear *stepwise* conclui-se que todas as variáveis foram consideradas significativas, com um p-value <0,001. Isto acontece porque a dimensão da amostra é mesmo muito elevada (Lin et al., 2013). A variável explicativa VLR é a que detém um maior coeficiente de determinação parcial (0,7978), tendo um peso maior na variância total explicada pelo modelo.

Com a exclusão das variáveis VL, VB, nr_trx e dl_tot, o valor de R_a^2 passou de 80,13% para 80,11%, mas os valores de VIF diminuíram drasticamente, permanecendo apenas a variável VLR com um VIF superior a 5 (Tabela 13). Isto significa que esta variável pode ser prevista como combinação linear das restantes variáveis independentes. Esta variável é muito importante na previsão do CLV e por esta razão excluiu-se a hipótese de a eliminar do modelo. Assim, é necessário analisar as restantes variáveis e identificar qual delas é que apresenta uma elevada correlação com ela.

Linear Regression Results

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: vlr_resp

Number of Observations Read	3389828
Number of Observations Used	3389828

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	20	3.363355E12	1.681678E11	682682	<.0001
Error	3.39E6	8.350245E11	246334		
Corrected Total	3.39E6	4.19838E12			

Root MSE	496.32041	R-Square	0.8011
Dependent Mean	858.71384	Adj R-Sq	0.8011
Coeff Var	57.79811		

Figura 13 Regressão Linear com as variáveis independentes - ANOVA e Principais medidas de desempenho do modelo

Tabela 13 Regressão Linear com as variáveis independentes - Parâmetros estimados, Teste de significância dos coeficientes, coeficientes estimados estandardizados, Tolerância e VIF

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	Intercept	1	177.13961	1.51955	116.57	<.0001	0	.	0
VLR	VLR	1	0.91655	0.00058029	1579.45	<.0001	0.88577	0.18656	5.36027
DESCONTOS	DESCONTOS	1	0.50002	0.00462	108.23	<.0001	0.04773	0.30167	3.31488
t_recency		1	-0.61648	0.00447	-138.02	<.0001	-0.03953	0.71534	1.39793
perc_sem		1	-98.27525	1.75297	-56.06	<.0001	-0.02658	0.26097	3.83187
cesta_rep		1	-0.67145	0.01528	-43.93	<.0001	-0.01277	0.69455	1.43978
primav		1	194.96198	3.12526	62.38	<.0001	0.01533	0.97188	1.02894
segm_baby_junior		1	25.90971	0.75564	34.29	<.0001	0.00948	0.76819	1.30176
segm_grocer		1	-77.07071	6.98071	-11.04	<.0001	-0.00271	0.97662	1.02394
cart_uni		1	99.53128	1.24479	79.96	<.0001	0.02083	0.86441	1.15686
DL_SP		1	-0.05514	0.00226	-24.45	<.0001	-0.00907	0.42620	2.34631
vb_eco		1	0.02114	0.00132	16.00	<.0001	0.00462	0.70252	1.42345
nr_ins		1	-1.50252	0.30606	-4.91	<.0001	-0.00166	0.51485	1.94232
vb_nalim		1	-0.14100	0.00199	-70.86	<.0001	-0.02812	0.37254	2.68430
FLAG_CLOSER_PARTNER		1	38.41986	0.77251	49.73	<.0001	0.01235	0.95204	1.05037
perc_vb_eco		1	22.29067	2.17752	10.24	<.0001	0.00318	0.60935	1.64109
idade_Median		1	-0.00144	0.00005062	-28.44	<.0001	-0.00742	0.86276	1.15907
agregado_Median		1	2.26404	0.12513	18.09	<.0001	0.00440	0.99073	1.00936
lifetime_Median		1	-0.01120	0.00027383	-40.91	<.0001	-0.01084	0.83570	1.19660
osd_Median		1	-5.67743	1.13612	-5.00	<.0001	-0.00123	0.97348	1.02724
perc_vb_na_Median		1	-11.94941	1.78474	-6.70	<.0001	-0.00187	0.75128	1.33106

Na Tabela 13, a variável que apresenta um VIF mais próximo de 5 é a variável perc_sem. No subcapítulo 3.1.1 Análise Exploratória dos Dados, na Tabela 8 Matriz de Correlações entre as variáveis quantitativas selecionadas é possível observar que a variável perc_sem apresenta a segunda maior correlação com a variável VLR (0,74), abaixo da correlação de 0,79 com a variável descontos. Testou-se a exclusão de cada uma das variáveis e o impacto nas medidas de desempenho é inferior quando se excluiu a variável perc_sem (R_a^2=80,09%) relativamente à exclusão da variável descontos (R_a^2=80,04%). Neste processo, a variável perc_vb_na_median foi excluída do modelo porque não apresentava um coeficiente relevante ($\beta_{perc\_vb\_na\_median} \approx 0$). Segundo os valores apresentados de VIF o problema da multicolinearidade estaria resolvido, mas os valores *condition index* contrariam esta conclusão, com valores superiores a 10 (Tabela 14). As proporções de variância correspondentes à décima nona dimensão onde o *condition index* é igual a 17,57 são superiores a 0,05 no coeficiente de regressão correspondente à ordenada na origem (ver Anexo 1).

Tabela 14 Regressão Linear com variáveis independentes, sem perc_sem - Diagnóstico de Colinearidade

Number	Eigenvalue	Condition Index
1	8.74205	1.00000
2	1.83429	2.18309
3	1.22835	2.66776
4	1.02718	2.91732
5	1.00067	2.95571
6	0.92653	3.07169
7	0.84426	3.21787
8	0.63593	3.70768
9	0.56791	3.92343
10	0.41597	4.58431
11	0.36921	4.86596
12	0.34447	5.03770
13	0.30466	5.35671
14	0.22035	6.29870
15	0.16524	7.27366
16	0.14174	7.85331
17	0.11851	8.58857
18	0.08435	10.18011
19	0.02831	17.57113

Segundo João Maroco (2007), quando isto acontece significa que alguma ou algumas das variáveis independentes consideradas possuem uma dispersão reduzida que torna essa

variável quase colinear com o vetor de 1's utilizado para estimação do coeficiente da ordenada na origem. Este fenómeno denomina-se por colinearidade não-essencial. A solução é estandardizar todas as variáveis de forma a ficarem com a mesma média e uma dispersão comparável. A estandardização também permite que as diferentes amplitudes de cada variável não tenham impacto na previsão.

Observando os resultados da regressão, com as variáveis estandardizadas, já se pode afirmar que não se verifica a existência de multicolinearidade entre as variáveis consideradas no modelo.

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: vlr_resp

Number of Observations Read	3389828
Number of Observations Used	3389828

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	18	2714992	150833	757659	<.0001
Error	3.39E6	674835	0.19908		
Corrected Total	3.39E6	3389827			

Root MSE	0.44618	R-Square	0.8009
Dependent Mean	-5.3188E-13	Adj R-Sq	0.8009
Coeff Var	-8.38878E13		

Figura 14 Regressão Linear com as variáveis estandardizadas - ANOVA e Principais medidas de desempenho do modelo

Tabela 15 Regressão Linear com as variáveis estandardizadas - Parâmetros estimados, Teste de significância dos coeficientes, coeficientes estimados estandardizados, Tolerância e VIF

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-6.3466E-13	0.00024234	-0.00	1.0000	0	.	0	-0.00047497	0.00047497
VLR	1	0.87149	0.00048662	1790.92	<.0001	0.87149	0.24801	4.03207	0.87053	0.87244
DESCONTOS	1	0.04687	0.00043988	106.56	<.0001	0.04687	0.30352	3.29469	0.04601	0.04774
t_recency	1	-0.03380	0.00026769	-126.28	<.0001	-0.03380	0.81958	1.22014	-0.03433	-0.03328
cesta_rep	1	-0.00605	0.00026486	-22.84	<.0001	-0.00605	0.83714	1.19455	-0.00657	-0.00553
primav	1	0.01500	0.00024574	61.04	<.0001	0.01500	0.97249	1.02829	0.01452	0.01548
segm_baby_junior	1	0.00910	0.00027549	33.01	<.0001	0.00910	0.77378	1.29235	0.00856	0.00964
segm_grocer	1	-0.00261	0.00024352	-10.74	<.0001	-0.00261	0.99033	1.00977	-0.00309	-0.00214
cart_uni	1	0.02019	0.00026039	77.55	<.0001	0.02019	0.86618	1.15449	0.01968	0.02070
DL_SP	1	-0.01077	0.00036997	-29.10	<.0001	-0.01077	0.42905	2.33071	-0.01149	-0.01004
vb_eco	1	0.00371	0.00028846	12.85	<.0001	0.00371	0.70580	1.41684	0.00314	0.00427
nr_ins	1	-0.00671	0.00032504	-20.65	<.0001	-0.00671	0.55587	1.79898	-0.00735	-0.00607
vb_nalim	1	-0.02806	0.00036184	-77.56	<.0001	-0.02806	0.44855	2.22942	-0.02877	-0.02735
FLAG_CLOSER_PARTNER	1	0.01081	0.00024673	43.82	<.0001	0.01081	0.96473	1.03656	0.01033	0.01130
perc_vb_eco	1	0.00722	0.00030025	24.05	<.0001	0.00722	0.65147	1.53500	0.00663	0.00781
idade_Median	1	-0.00764	0.00026040	-29.33	<.0001	-0.00764	0.86608	1.15463	-0.00815	-0.00713
agregado_Median	1	0.00471	0.00024340	19.37	<.0001	0.00471	0.99126	1.00882	0.00424	0.00519
lifetime_Median	1	-0.01273	0.00026287	-48.41	<.0001	-0.01273	0.84990	1.17661	-0.01324	-0.01221
osd_Median	1	-0.00232	0.00024484	-9.46	<.0001	-0.00232	0.97964	1.02078	-0.00280	-0.00184

Tabela 16 Regressão Linear com variáveis estandardizadas - Diagnóstico de Colinearidade (Tabela completa em Anexo 1)

Number	Eigenvalue	Condition Index
1	3.89554	1.00000
2	1.85196	1.45033
3	1.21478	1.79075
4	1.12462	1.86115
5	1.07864	1.90040
6	1.04461	1.93111
7	1.00000	1.97371
8	0.99740	1.97628
9	0.96969	2.00432
10	0.94424	2.03116
11	0.90228	2.07784
12	0.77038	2.24870
13	0.71530	2.33367
14	0.62663	2.49332
15	0.55158	2.65755
16	0.49705	2.79953
17	0.39897	3.12475
18	0.23030	4.11277
19	0.18606	4.57571

A próxima etapa é a análise dos resíduos e da sua distribuição. Na avaliação dos pressupostos recorreu-se à representação gráfica em alternativa aos testes de hipóteses, não adequados para este conjunto de dados de tão elevada dimensão. Um dos gráficos importantes a analisar é a distribuição dos resíduos da regressão pelos diferentes valores

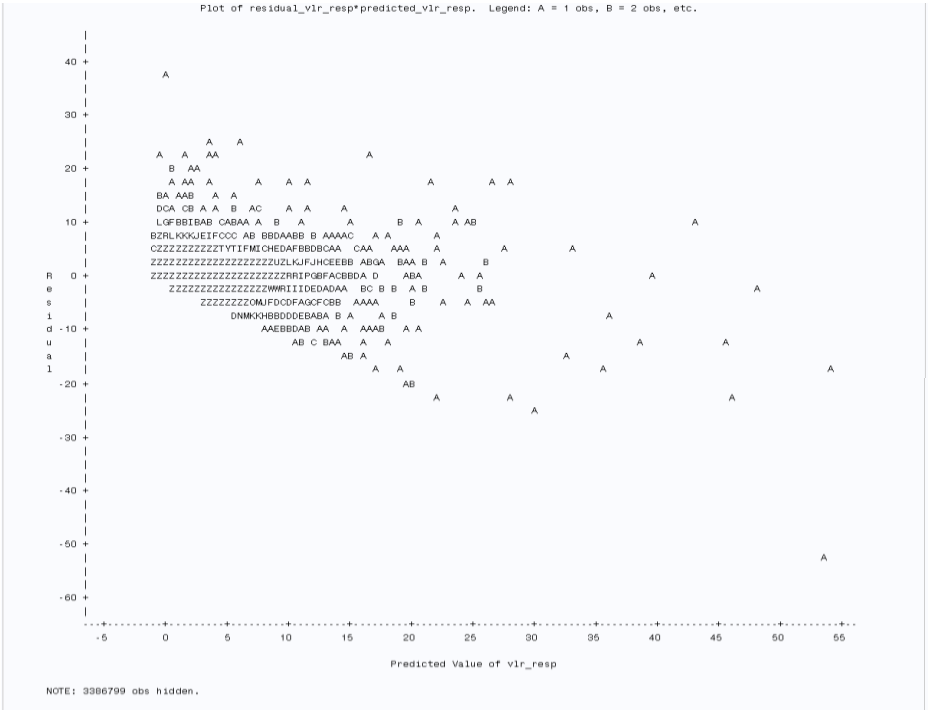


Figura 15 Diagrama de dispersão dos resíduos com a variável resposta estimada

estimados da variável resposta de modo a avaliar se os resíduos têm a mesma dispersão ao longo dos diferentes valores tomados pela variável resposta (Figura 15).

O diagrama de dispersão acima apresentado permite retirar duas conclusões: os resíduos aumentam a sua dispersão quanto maior for o valor de $\widehat{CLV}$. Ou seja, os resíduos não possuem uma variância constante. A segunda conclusão retirada é a existência de *outliers* na base de dados. Estes *outliers* devem ser eliminados da base de dados porque estão a influenciar severamente a estimação da variável resposta. É possível identificar estes valores com recurso aos valores das medidas *DfFits* e Distância de *Cook* (Maroco, 2007).

Para uma amostra grande, os valores de *DfFits* e Distância de *Cook* superiores a $2\sqrt{p/n} \approx 0{,}0049$ (Christensen, 1997) e a $4/n \approx 0{,}000001180$, respetivamente, devem ser investigados. Como é possível observar na Tabela 17, ambas as medidas apresentam valores máximos superiores ao limite definido como aceitável.

Tabela 17 Medidas de centralidade e dispersão das variáveis *DfFits* e Distância de *Cook*

Variable	Label	Mean	Std Dev	Minimum	Maximum	N
cookd_vlr_resp	Cook's D Influence Statistic	0.000	0.020	0.000	36.765	3.389.828
dffits_vlr_resp	Standard Influence on Predicted Value	-0.000	0.017	-26.487	1.940	3.389.828

Na base de dados excluíram-se as observações que apresentaram uma Distância de *Cook* superior a 0,000001180, que eram cerca de 181.015 observações.

Os resultados de desempenho do modelo sem estas observações estão representados na Figura 16. O valor de R_a^2 aumentou de 80.09% para 86,75% e o erro quadrático médio diminuiu de 0,44618 para 0.29529.

Number of Observations Read	3208813
Number of Observations Used	3208813

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	1831918	107760	1235831	<.0001
Error	3.21E6	279795	0.08720		
Corrected Total	3.21E6	2111712			

Root MSE	0.29529	R-Square	0.8675
Dependent Mean	-0.10000	Adj R-Sq	0.8675
Coeff Var	-295.27752		

Figura 16 Regressão Linear com variáveis independentes e estandardizadas sem *outliers* segundo a Distância de *Cook*

Apenas um dos problemas identificados está resolvido. A variância dos erros continua inconstante como se pode verificar através da Figura 17.

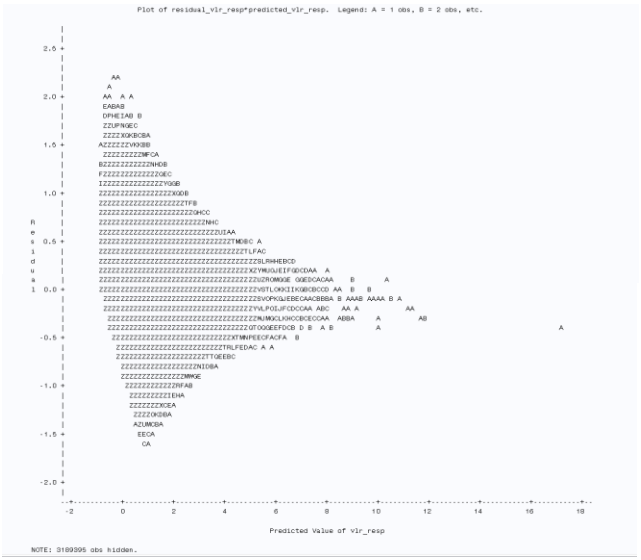


Figura 17 Regressão Linear com variáveis independentes e estandardizadas sem *outliers* segundo a Distância de *Cook*: Diagrama de dispersão dos resíduos versus $\widehat{CLV}$

Para contornar o problema da variância inconstante dos erros, neste caso, deve-se fazer uma transformação na variável resposta, usando o logaritmo da variável em vez do valor absoluto da mesma na previsão do CLV. É importante referir que, para evitar problemas com valores nulos, a transformação aplicada foi $vlr\_resp = \log(vlr\_resp + 1)$. Na Figura 18, com esta transformação, o diagrama de dispersão dos erros versus a variável resposta estimada já indicia uma variância constante. Também é possível observar que ainda existem observações que são outliers na base de dados e que devem ser estudadas.

Recorrendo novamente à Distância de *Cook* identificaram-se os outliers. Uma vez que a dimensão da base de dados nesta fase é composta por 3.208.813 clientes, o corte utilizado para definir se um cliente é ou não é um *outlier* tem o valor de 0,000001247. Com isto, mais 121.149 clientes foram identificados como *outliers* e foram excluídos da análise. Assim, os resíduos apresentam agora uma variância contante como se pode constatar através da Figura 19.

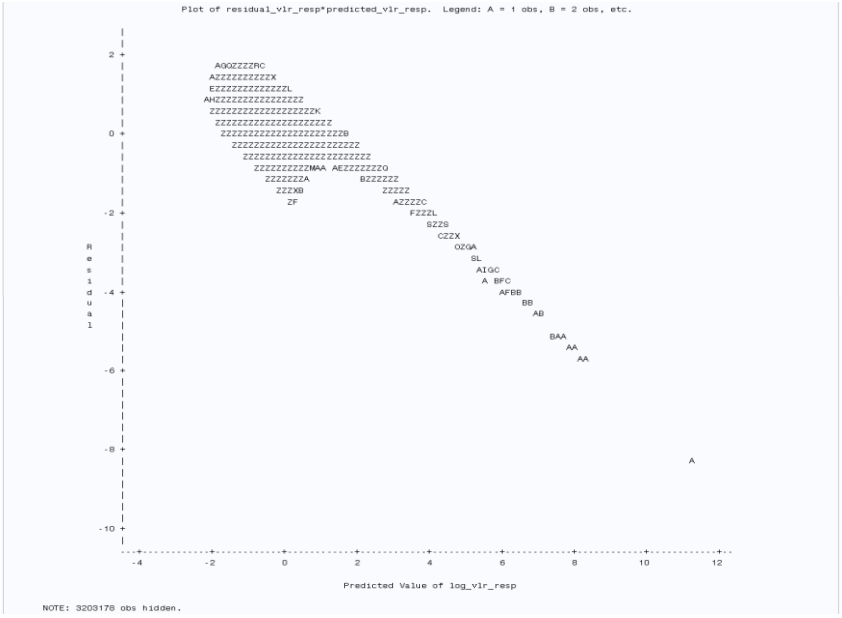


Figura 18 Regressão Linear com variáveis independentes e estandardizadas sem outliers segundo a Distância de Cook: Diagrama de dispersão dos resíduos versus Logaritmo de ($\widehat{CLV}$)

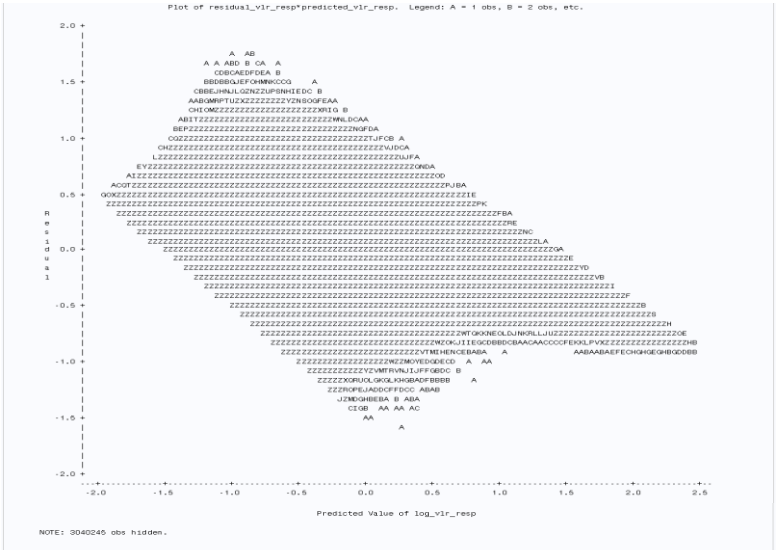


Figura 19 Diagrama de dispersão após logaritmação da variável resposta e exclusão de *outliers*

De modo a avaliar se os resíduos se distribuem segundo uma distribuição normal com média igual a 0 e variância unitária, produziu-se o histograma e o gráfico Q-Q dos resíduos do modelo. Uma vez que o histograma apresenta uma curva muito semelhante à distribuição normal e os pontos no gráfico Q-Q estão sobrepostos à diagonal dos quantis de uma distribuição normal pode-se concluir que o pressuposto da normalidade dos resíduos se cumpre.

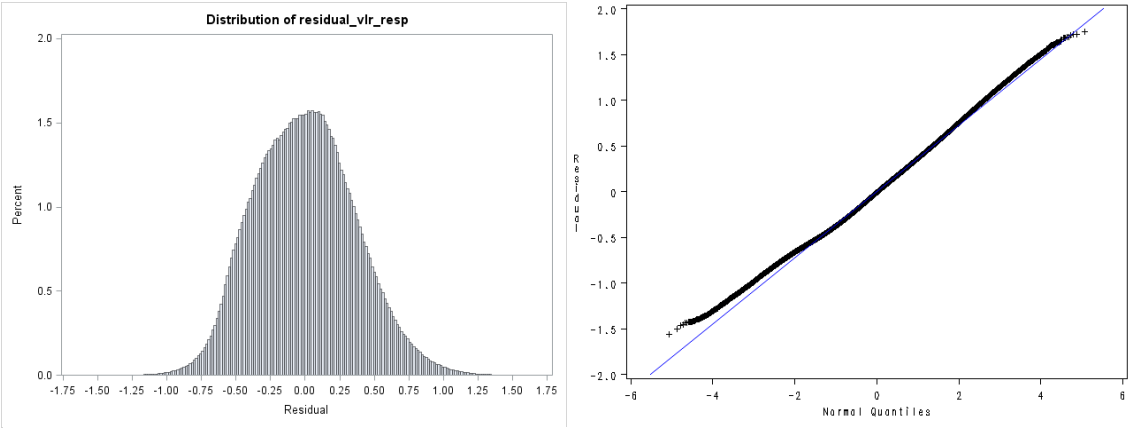


Figura 20 Histograma e Gráfico Q-Q dos resíduos do modelo

Em suma, todos os pressupostos da aplicação da Regressão Linear são cumpridos neste modelo. O *output* completo da regressão linear é apresentado abaixo na Figura 21, juntamente com a seleção de variáveis *stepwise*.

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	VLR		VLR	1	0.7213	0.7213	729446	7989606	<.0001
2	t_recency			2	0.0433	0.7645	137098	567168	<.0001
3	nr_ins			3	0.0030	0.7675	96619.1	39252.6	<.0001
4	perc_vb_eco			4	0.0028	0.7702	58797.3	37117.1	<.0001
5	DESCONTOS		DESCONTOS	5	0.0014	0.7716	40170.0	18390.0	<.0001
6	FLAG_CLOSER_PARTNER			6	0.0005	0.7721	33026.1	7070.30	<.0001
7	cart_uni			7	0.0005	0.7726	25973.2	6996.04	<.0001
8	DL_SP			8	0.0004	0.7731	19948.1	5988.50	<.0001
9	cesta_rep			9	0.0004	0.7735	14296.0	5628.03	<.0001
10	vb_eco			10	0.0004	0.7738	9390.67	4892.47	<.0001
11	lifetime_Median			11	0.0003	0.7741	5327.51	4058.18	<.0001
12	vb_nalim			12	0.0002	0.7743	3070.56	2256.72	<.0001
13	segm_baby_junior			13	0.0001	0.7744	1319.67	1752.15	<.0001
14	idade_Median			14	0.0001	0.7745	525.522	796.01	<.0001
15	segm_grocer			15	0.0000	0.7745	157.061	370.44	<.0001
16	primav			16	0.0000	0.7745	69.0250	90.03	<.0001
17	agregado_Median			17	0.0000	0.7745	22.0547	48.97	<.0001
18	osd_Median			18	0.0000	0.7745	19.0000	5.05	0.0246
19		osd_Median		17	0.0000	0.7745	22.0547	5.05	0.0246

Number of Observations Read	3087664
Number of Observations Used	3087664

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	1391877	81875	623915	<.0001
Error	3.09E6	405186	0.13123		
Corrected Total	3.09E6	1797063			

Root MSE	0.36225	R-Square	0.7745
Dependent Mean	-0.44873	Adj R-Sq	0.7745
Coeff Var	-80.72833		

Figura 21 *Output* completo da Regressão Linear para o total da população

Parameter Estimates											
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	-0.34696	0.00021231	-1634.2	<.0001	0	.	0	-0.34738	-0.34655
VLR	VLR	1	0.66783	0.00058070	1150.05	<.0001	0.64891	0.22936	4.35988	0.66670	0.66897
DESCONTOS	DESCONTOS	1	0.05400	0.00048863	110.51	<.0001	0.05448	0.30047	3.32811	0.05304	0.05496
t_recency		1	-0.15496	0.00022857	-677.96	<.0001	-0.20432	0.80399	1.24380	-0.15541	-0.15451
cesta_rep		1	0.02012	0.00027139	74.13	<.0001	0.02190	0.83655	1.19538	0.01959	0.02065
primav		1	0.00382	0.00040354	9.46	<.0001	0.00257	0.99096	1.00912	0.00303	0.00461
segm_baby_junior		1	0.00910	0.00024299	37.44	<.0001	0.01158	0.76311	1.31042	0.00862	0.00957
segm_grocer		1	-0.00782	0.00040679	-19.24	<.0001	-0.00520	0.99867	1.00133	-0.00862	-0.00703
cart_uni		1	0.02558	0.00024482	104.48	<.0001	0.03051	0.85662	1.16738	0.02510	0.02606
DL_SP		1	0.03055	0.00043128	70.83	<.0001	0.03066	0.38978	2.56553	0.02970	0.03139
vb_eco		1	0.03141	0.00044925	69.91	<.0001	0.02637	0.51344	1.94765	0.03053	0.03229
nr_ins		1	0.06283	0.00029075	216.09	<.0001	0.08060	0.52489	1.90517	0.06226	0.06340
vb_nalim		1	-0.02479	0.00043076	-57.55	<.0001	-0.02297	0.45828	2.18205	-0.02563	-0.02394
FLAG_CLOSER_PARTNER		1	0.02016	0.00021531	93.65	<.0001	0.02581	0.96153	1.04001	0.01974	0.02059
perc_vb_eco		1	-0.05466	0.00028066	-194.74	<.0001	-0.07094	0.55029	1.81721	-0.05521	-0.05411
idade_Median		1	-0.00627	0.00022117	-28.35	<.0001	-0.00824	0.86379	1.15769	-0.00670	-0.00584
agregado_Median		1	-0.00161	0.00023054	-7.00	<.0001	-0.00190	0.99195	1.00812	-0.00207	-0.00116
lifetime_Median		1	0.01547	0.00022558	68.58	<.0001	0.02013	0.84798	1.17928	0.01503	0.01591

Figura 22 Output completo da Regressão Linear para o total da população (cont.)

O modelo final escreve-se então:

$$\begin{aligned}\widehat{vlr\_resp} = -0{,}347 + {} & 0{,}668VLR + 0{,}054descontos - 0{,}155t_{recency} \\ & + 0{,}020cesta_{rep} + 0{,}004\text{primav} + 0{,}009\text{segm\_baby\_junior} \\ & - 0{,}008\text{segm\_grocer} + 0{,}026\text{cart\_uni} + 0{,}031\text{DL\_SP} \\ & + 0{,}031\text{vb\_eco} + 0{,}063\text{nr\_ins} - 0{,}025\text{vb\_nalim} \\ & + 0{,}02\text{FLAG\_CLOSER\_PARTNER} - 0{,}055\text{perc\_vb\_eco} \\ & - 0{,}006\text{idade\_Median} - 0{,}002\text{agregado\_Median} \\ & + 0{,}015\text{lifetime\_Median}\end{aligned}$$

Todos os coeficientes da regressão são significativos (embora os p-values não possam ser interpretados devido à dimensão do conjunto em análise (Lin et al., 2013)) e o modelo ajustado explica 77,45% da variância de CLV e com um erro absoluto médio de 492,15. Não há problemas de multicolinearidade. Segundo os valores dos coeficientes de cada variável na equação, as variáveis que contribuem mais para a previsão do CLV são as VLR e t_recency, como já se tinha verificado através do coeficiente de determinação parcial nos resultados da regressão *stepwise*. As variáveis que menos contribuem são a idade_median, primav e agregado_median.

Tendo por base o modelo acima descrito é possível concluir que o CLV do cliente será maior (Figura 23):

- quanto maior for o valor das suas vendas líquidas reportadas no ano anterior, pois se o cliente já normalmente gasta um determinado valor na empresa espera-se que esse valor se mantenha;
- quanto maior for o nível de descontos, sejam eles em cartão, no ecossistema ou diretos, ou seja, quantas mais ofertas o cliente tiver, maior será o valor de desconto inerente às suas compras e maior será o valor gasto;
- quanto menor for o intervalo de tempo entre a última visita e a última data do ano em estudo, pois se quanto mais recente for a compra do cliente maior a probabilidade de o cliente voltar e por consequência elevar o seu gasto total;
- quanto maior for o gasto médio por transação, pois tendencialmente esse comportamento se irá manter, e no total ano este cliente gastará mais;
- se tiver filhos, o que faz todo o sentido, uma vez que esta facto está aliado à necessidade de um maior investimento na alimentação;
- se não for um comerciante, pois ao contrário do que se suspeitava, este cliente tem um grande valor de vendas, mas no total ano não é tão elevado como um cliente fiel da empresa;
- quanto maior for o número de insígnias no ecossistema visitadas pelo cliente e o valor que gastam nessas insígnias, indicando um grande envolvimento com o ecossistema, e consequente com a empresa em estudo; Mas se o peso de vendas no ecossistema for muito superior ao peso de vendas efetuadas na área de retalho, indicia um baixo CLV de cliente;
- se a uma loja da empresa estiver mais próxima da residência desse cliente do que uma loja da concorrência;
- quanto menor for a idade e o agregado familiar do cliente; A idade pode dever-se ao facto de quanto maior a idade menor é a mobilidade do cliente, substituindo por uma loja mais perto ou as suas compras serem realizadas por outra pessoa; Ao contrário do que se esperava o agregado tem um impacto negativo na previsão do CLV; Isto acontece talvez porque o agregado familiar estão por vezes associados a classes sociais mais baixas e, por consequência, a um estilo de vida mais económico;
- quanto maior for o tempo que o cliente se encontra associado à empresa.

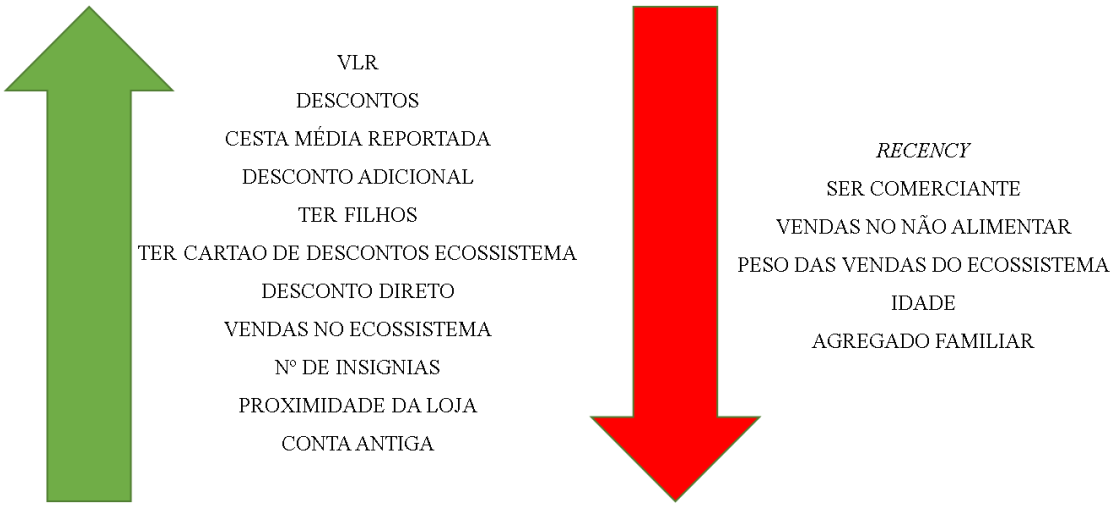


Figura 23 Contribuição de cada variável para o aumento do CLV do cliente

Uma vez que as variáveis foram estandardizadas e a variável resposta foi logaritmizada, os valores dos coeficientes apresentados na equação acima têm a seguinte relação com a os coeficientes reais:

$$\beta_j' = \exp(\beta_j \frac{s_{x_j}}{s_{vlr_{resp}}})$$

em que β_j' são os coeficientes estimados do modelo com variáveis estandardizadas, β_j são os coeficientes reais, s_{x_j} são os desvios-padrão das variáveis explicativas X_j e s_{vlr_resp} é o desvio padrão da variável resposta vlr_resp.

É importante referir que o erro aqui obtido é um indicador do desempenho do modelo usando o conjunto de treino para validação do modelo. A isto se chama avaliação por resubstituição. Este método não é o mais correto pois produz estimativas otimistas, levando à sobreavaliação do modelo. Existem vários métodos de avaliação do desempenho de um modelo preditivo que consistem no uso de amostras distintas para treino e teste do modelo. A mais simples denomina-se de *Holdout* e consiste na divisão do conjunto a analisar em duas amostras, geralmente nas proporções de 70% e 30%, para treino e validação do modelo, respetivamente. Outro método mais robusto é a Validação Cruzada que divide o conjunto em *r* subamostras e usa *r* – 1 para treino e a subamostra restante é utilizada na validação do modelo. Esse processo é repetido r vezes, utilizando em cada ciclo uma sub-amostra diferente para teste (Gama, Carvalho, Faceli, Lorena, &

Oliveira, 2015). A Validação Cruzada sendo um processo que requer muito esforço computacional e é mais indicado para amostras mais pequenas, o método aplicado para medir o desempenho do modelo será o *Holdout*.

Validação do modelo da Regressão Linear para a população- Método *Holdout*

O tipo de amostragem aplicado na base de dados foi amostragem aleatória estratificada que consiste em retirar uma amostra aleatoriamente da população, sem reposição, mas garantindo que a amostra resultante e a população terão a mesma proporção de uma determinada classe. Este tipo de amostragem foi o selecionado na aplicação porque os clientes possuem diferentes segmentos valor (os mais leais, os frequentes, os ocasionais, e os sem valor) e é importante prever o comportamento dos quatro segmentos de clientes existentes, e esta amostragem garante que a proporção destes clientes na amostra é muito próxima da proporção existente na população.

A amostra de treino é composta por 2.372.880 clientes (70% da população) e amostra de teste é composta pelos restantes 1.016.948 clientes (30%). Para medir o erro calcular-se-á o erro quadrático médio, o erro absoluto médio e o coeficiente de correlação de *Spearman* entre o CLV real e o CLV estimado. Antes do cálculo dos erros foi necessário converter os valores de Log CLV' e o $\widehat{CLV'}$ para os valores reais. A relação que cada uma delas tem com o seu valor real é a seguinte:

$$CLV = (\exp(CLV') - 1) * s_{CLV} + \overline{x_{CLV}},$$

CLV corresponde ao valor real, o CLV' é o valor da variável com as transformações aplicadas acima (estandardizada e logaritmizada), s_{CLV} é o desvio padrão de CLV e $\overline{x_{CLV}}$ é a respetiva média.

Pela avaliação por resubstituição o valor do RMSE foi de 492,15, de MAE foi de 280,13 e o coeficiente de correlação de *Spearman* tinha o valor de 0,878, que corresponde a uma correlação elevada entre a variável resposta e a respetiva previsão.

Através do método de *holdout*, o modelo de Regressão Linear apresenta valores menos otimistas que os apresentados pela resubstituição, sendo também influenciados negativamente pela inclusão das observações *outliers* excluídas anteriormente. Aplicando

as mesmas fórmulas do RMSE e MAE utilizadas por Nicolas Glady e restantes autores em 2009, obtêm-se os valores dos erros representados na Tabela 18:

Tabela 18 RMSE e MAE e respetivos desvios padrão, com 99% das melhores previsões com método *Holdout*

m_mae	std_mae	m_rmse	std_rmse
1036,43046	2699,089351	2891,239213	7238,660588

De modo a identificar onde o modelo concentra o seu erro, efetuaram-se vários cortes: por segmento valor (Tabela 19) e por grupos homogéneos de clientes segundo o valor de CLV correspondente (Tabela 20).

Tabela 19 RMSE e MAE e respetivos desvios padrão por Segmento Valor

segm_valor	m_mae	std_mae	m_rmse	std_rmse	Nº Clientes
Loyal	3.313,02	5.022,57	6.016,83	10.352,91	198.651
Frequent	624,12	1.247,10	1.394,55	4.524,57	430.710
Occasional	225,21	375,77	438,09	1.631,90	348.778
Sem Valor	1.325,38	3.067,25	3.341,31	7.760,34	28.640

O erro é inferior em clientes do segmento *Occasional*, com um erro absoluto médio de 225, 21€ e um erro quadrático médio de 438,09€, muito inferior aos erros observados para os restantes segmentos. Com isto, pode-se afirmar que o modelo prevê melhor o CLV para clientes que não são tão fieis à empresa em estudo.

Os grupos de clientes criados para estudar os valores do erro têm por base cortes segundo os quartis da variável resposta. Ou seja, o primeiro grupo é composto pelos clientes que apresentam valores de CLV inferiores ao valor do primeiro quartil do CLV; o segundo grupo de clientes apresentam valores entre o primeiro e segundo quartil; o terceiro entre o segundo e o terceiro; e o quarto apresenta valores de CLV superiores ao valor do terceiro quartil.

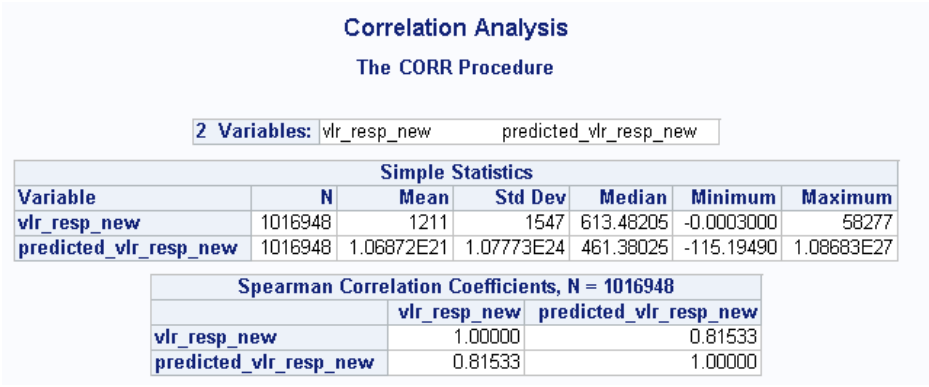
Tabela 20 RMSE e MAE e respetivos desvios padrão por Grupos Homogéneos

Legenda: G1: vlr_resp <= Q1;
G2: Q1 < vlr_resp <= Q2;
G3: Q2 < vlr_resp <= Q3;
G4: vlr_resp > Q3;

GP Predicted vlr_resp

	Sum of nr_id	G1	G2	G3	G4	Grand Total
GP vlr_resp	G1	34,7%	57,4%	6,5%	1,3%	100,0%
	G2	8,8%	72,1%	16,2%	2,9%	100,0%
	G3	3,5%	36,7%	48,8%	11,0%	100,0%
	G4	0,4%	8,0%	24,3%	67,3%	100,0%

Assim, tem-se que 72% dos clientes do grupo G2 estão bem classificados. Para os clientes com um CLV inferior (G1), a taxa de acerto é muito baixa, no valor de 34,7%, e a maioria destes clientes estão classificados como pertencentes ao G2, grupo adjacente ao G1. Os clientes do G4 também apresentam uma taxa de acerto elevada, de 67,3%. Esta abordagem permite perceber se a previsão se encontra, pelo menos, na mesma ordem de grandeza que o valor real.



Correlation Analysis

The CORR Procedure

2 Variables: vlr_resp_new predicted_vlr_resp_new

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
vlr_resp_new	1016948	1211	1547	613.48205	-0.0003000	58277
predicted_vlr_resp_new	1016948	1.06872E21	1.07773E24	461.38025	-115.19490	1.08683E27

Spearman Correlation Coefficients, N = 1016948

	vlr_resp_new	predicted_vlr_resp_new
vlr_resp_new	1.00000	0.81533
predicted_vlr_resp_new	0.81533	1.00000

Figura 24 Coeficiente de *Spearman* entre o CLV e o CLV estimado

Na mesma ótica, o coeficiente de *Spearman* permite-nos também chegar a essa conclusão, uma vez que se trata da medição da correlação entre os *ranks* do CLV e os *ranks* do CLV estimado. O coeficiente de correlação de *Spearman* entre a variável resposta e o seu valor previsto é de 0,8153, um valor muito superior a 0,5, indicando assim uma forte correlação positiva entre as duas variáveis.

Através do Diagrama de Dispersão entre CLV e o $\widehat{CLV}$ (Figura 25) conclui-se que a as previsões do modelo são sobrestimadas, prevendo que o cliente gastará mais do que ele efetivamente gasta.

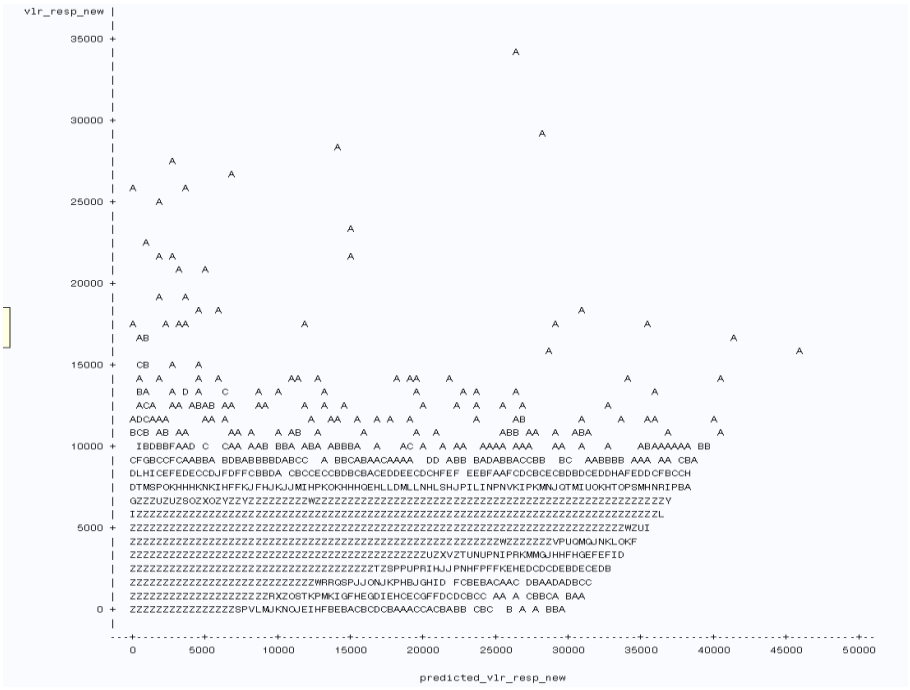


Figura 25 Diagrama de dispersão entre o CLV e o $\widehat{CLV}$

4.1.2 Regressão Linear aplicada a amostras estratificadas

Após o desenvolvimento do modelo com o conjunto total dos dados, ele será utilizado para prever o CLV para as 100 amostras de 1000 clientes retiradas do conjunto total.

A empresa possui um elevado número de clientes, na ordem dos 3 Milhões. Trabalhar com uma base desta dimensão impede o uso de testes de hipóteses e requer um grande esforço computacional. Determinados *softwares* ou modelos não estão preparados para serem aplicados a um número tão elevado de observações. O Modelo de Pareto/NBD apresenta esta limitação. O *package* BTYD, construído para aplicação do Modelo de Pareto/NBD, não suporta base de dados com mais de 26.2 Mb. Neste contexto, a previsão CLV foi efetuada sobre várias amostras, nos três modelos.

O tipo de amostragem aplicado na base de dados foi a amostragem aleatória estratificada.

O erro quadrático médio e o erro absoluto médio obtidos em cada uma das 100 amostras estão representados na Figura 26 através de Box-Plots.

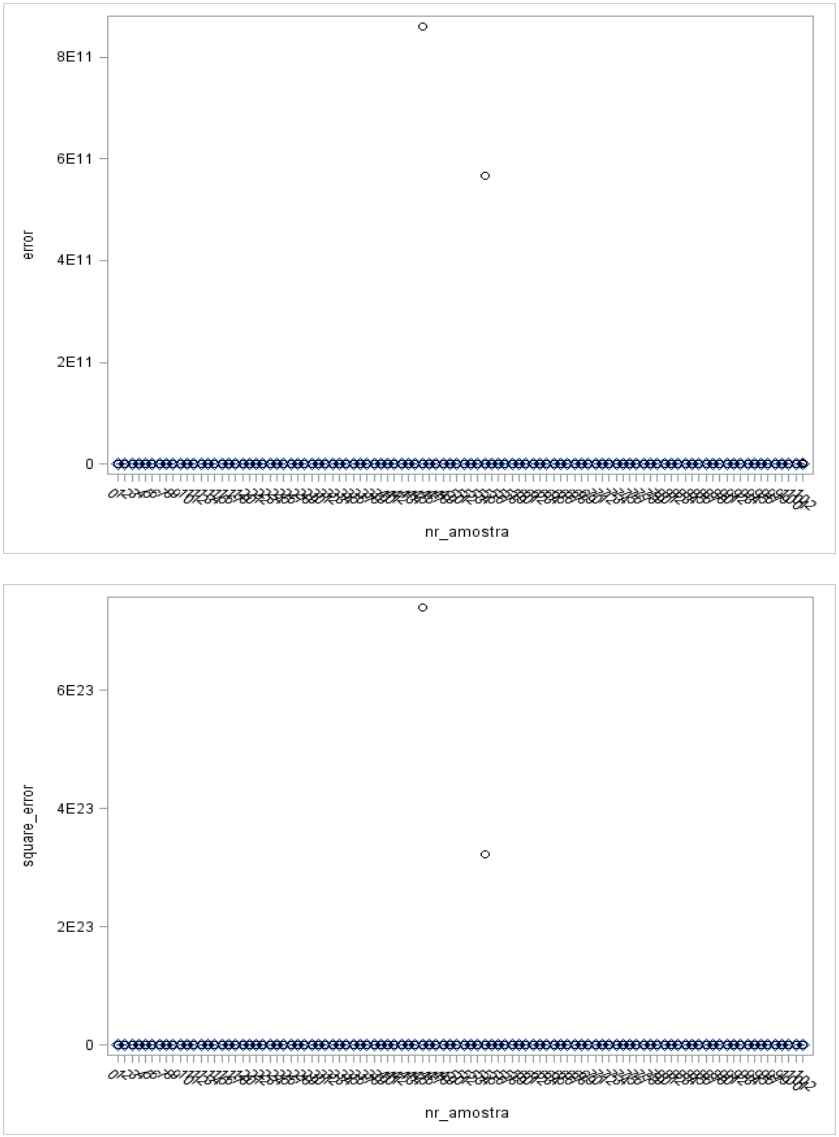


Figura 26 Box-plots dos valores do erro para as 100 amostras estratificadas

Pode-se verificar que existem duas observações que apresentam um erro muito elevado. Isto vem suportar a avaliação do erro avaliado com 99% das melhores previsões é mais acertado do que com o conjunto total de observações em análise. Aplicando RMSE e MAE com as melhores previsões obtém-se as previsões descritas na Tabela 21.

Tabela 21 RMSE e MAE e respetivos desvios-padrão, e o coeficiente de correlação de *Spearman* da Regressão Linear para as 100 amostras estratificadas

	error		square_error		Spearman		error		square_error		Spearman
amostra	Mean	Std Dev	Mean	Std Dev		amostra	Mean	Std Dev	Mean	Std Dev	
0	441,40	997,12	1 089,99	2 895,83	0,86	**51**	378,98	668,68	768,32	1 897,06	0,85
1	412,74	787,32	888,60	2 078,77	0,85	**52**	426,19	917,89	1 011,59	2 678,64	0,86
2	389,36	673,54	777,69	1 714,11	0,86	**53**	361,95	617,66	715,63	1 678,13	0,85
3	352,41	542,74	646,89	1 331,47	0,87	**54**	426,42	976,81	1 065,38	2 961,81	0,85
4	371,85	668,24	764,44	1 806,78	0,86	**55**	402,59	736,40	838,94	1 871,68	0,86
5	406,11	756,62	858,38	1 910,07	0,85	**56**	371,70	580,99	689,48	1 385,04	0,86
6	427,34	967,02	1 056,79	2 648,41	0,86	**57**	379,21	577,26	690,42	1 478,31	0,85
7	405,74	666,22	779,76	1 624,61	0,85	**58**	413,23	834,29	930,64	2 294,20	0,85
8	371,02	662,26	758,81	1 728,55	0,85	**59**	352,26	572,01	671,54	1 392,25	0,85
9	417,76	836,66	934,78	2 275,81	0,85	**60**	413,55	872,57	965,21	2 477,01	0,88
10	409,87	894,02	983,09	2 619,69	0,84	**61**	410,34	813,10	910,41	2 378,33	0,85
11	472,42	1 396,05	1 473,14	4 541,84	0,88	**62**	370,34	614,37	717,09	1 499,73	0,86
12	443,91	904,41	1 007,07	2 284,02	0,86	**63**	425,25	857,70	956,95	2 451,47	0,83
13	416,99	885,46	978,33	2 489,81	0,86	**64**	388,82	780,70	871,81	2 404,43	0,84
14	430,75	887,07	985,72	2 712,00	0,85	**65**	478,11	1 089,19	1 189,00	3 172,06	0,85
15	383,39	720,20	815,57	1 907,29	0,84	**66**	481,94	1 457,19	1 534,12	4 588,58	0,87
16	386,13	719,30	816,07	1 913,94	0,84	**67**	369,65	655,35	752,13	1 913,66	0,86
17	391,81	737,42	834,71	1 976,34	0,84	**68**	332,54	475,97	580,44	1 104,22	0,85
18	337,31	534,01	631,39	1 471,05	0,85	**69**	386,54	688,75	789,50	1 827,21	0,86
20	371,56	654,40	752,24	1 976,66	0,86	**70**	362,07	598,91	699,59	1 546,08	0,84
21	383,07	784,86	873,00	2 252,39	0,83	**71**	399,68	749,31	848,90	1 917,53	0,87
22	539,11	1 354,81	1 457,50	3 820,72	0,84	**72**	389,27	657,39	763,71	1 714,01	0,86
23	462,01	1 122,94	1 213,75	3 370,22	0,85	**73**	391,90	704,01	805,43	1 798,03	0,85
24	372,18	684,09	778,48	1 887,14	0,85	**74**	416,93	928,14	1 017,05	2 679,66	0,85
25	407,76	845,92	938,68	2 304,17	0,87	**75**	405,03	808,27	903,70	2 286,46	0,86
26	364,00	601,65	702,93	1 521,79	0,83	**76**	407,10	828,28	922,55	2 183,98	0,86
27	473,93	1 363,27	1 442,65	4 295,31	0,85	**77**	344,99	508,54	614,30	1 186,20	0,86
28	424,14	764,32	873,78	2 080,48	0,85	**78**	413,19	1 053,06	1 130,73	3 244,98	0,87
29	425,22	904,21	998,79	2 487,71	0,88	**79**	369,39	596,07	701,00	1 539,00	0,83
30	356,99	635,09	728,27	1 796,30	0,86	**80**	429,45	759,49	872,17	1 845,49	0,86
31	390,38	721,46	819,99	1 913,89	0,87	**81**	370,86	675,31	770,14	1 786,26	0,84
32	374,69	655,20	754,48	1 697,29	0,84	**82**	394,92	697,00	800,80	1 769,48	0,85
33	421,63	884,49	979,44	2 376,45	0,85	**83**	365,33	555,45	664,59	1 401,61	0,84
34	382,96	726,05	820,53	2 034,29	0,84	**84**	378,05	665,08	764,72	1 769,66	0,85
35	542,23	1 514,01	1 607,46	4 475,65	0,84	**85**	372,50	594,87	701,62	1 525,43	0,85
36	411,08	769,66	872,22	2 081,86	0,85	**86**	438,52	1 065,47	1 151,68	3 237,08	0,84
37	380,92	693,52	790,94	1 734,05	0,85	**87**	366,86	631,04	729,65	1 625,38	0,86
38	323,61	476,67	575,94	1 144,24	0,86	**88**	445,89	1 143,67	1 226,97	3 814,87	0,85
39	356,62	587,78	687,25	1 594,20	0,84	**90**	411,85	775,53	877,76	2 139,49	0,85
40	416,47	830,39	928,60	2 314,80	0,87	**91**	501,43	1 228,78	1 326,58	3 523,20	0,85
41	431,95	1 066,57	1 150,22	3 300,21	0,86	**92**	377,02	631,15	734,91	1 644,48	0,86
42	366,48	636,21	733,94	1 831,85	0,84	**93**	435,72	867,55	970,43	2 439,10	0,84
43	385,27	723,11	819,02	2 017,58	0,85	**94**	376,32	637,36	739,89	1 661,28	0,84
44	401,27	868,14	955,99	2 453,47	0,84	**95**	389,01	654,48	761,07	1 668,91	0,86
45	547,13	1 714,08	1 798,46	5 624,26	0,87	**96**	420,27	893,35	986,86	2 613,63	0,85
46	363,63	652,06	746,31	1 833,38	0,84	**97**	332,74	500,10	600,47	1 247,33	0,85
47	391,44	688,83	791,98	1 724,96	0,87	**99**	332,06	484,03	586,78	1 132,61	0,84
48	409,24	921,31	1 007,68	2 720,73	0,86	**100**	383,57	678,18	778,84	1 894,52	0,85
49	421,29	944,02	1 033,32	2 712,35	0,85	**101**	384,82	707,21	804,81	1 891,16	0,84
50	365,00	670,60	763,20	1 787,31	0,86	**102**	417,89	799,94	902,16	2 142,41	0,85

Os erros RMSE e MAE apresentam valores mais otimistas relativamente aos erros apresentados para a amostra total de teste. O coeficiente de correlação de Spearman apresenta valores acima de 80%, indicando um bom ajustamento do modelo às 100 amostras.

4.2 Modelo Pareto/NBD

No primeiro artigo que surge sobre o método de Pareto/NBD (Schmittlein et al., 1987), os autores procuram dar uma resposta a questões como: "quantos clientes temos?", "qual é a evolução dos nossos clientes?", "de todos os clientes quais deles são ativos no nosso negócio?" e "qual será o comportamento deles, individual e coletivamente, no próximo ano?". Assim, e no sentido de responder a estas questões, as variáveis necessárias para a previsão do CLV são o número de compras que o cliente tem no passado num determinado período no passado, qual a data da sua última compra e há quanto tempo o cliente está ativo na empresa.

O desenvolvimento deste modelo foi feito com recurso ao *software* R, através do *package* BTYD (Dziurzynski, Wadsworth, & McCarthy, 2015; McCarthy & Wadsworth, 2014). Este *package* possui funções para os dois submodelos que constituem o Modelo de Pareto/NBD: as funções dc e pnbd para a previsão do número de transações repetidas e spend para a previsão do gasto médio por transação.

No *package* BTYD a base de dados de entrada é composta pelo código de cliente, dia da compra e valor da mesma. Esta informação é trabalhada para ficar no formato: código de cliente, número de transações repetidas, tempo entre a primeira e última compra (t) e tempo entre a primeira compra e o último dia do período em análise (T_cohort). Todo este tratamento é feito dentro do *package*.

O *output* final é constituído pelos parâmetros de cada modelo, a previsão de número de transações e o valor média da transação por cliente. O CLV é obtido através da multiplicação do número de transações pelo valor médio de cada transação. O valor médio de cada transação é estimado apenas para clientes que repetiram a sua compra no passado. Assim, todos os clientes que estiverem nestas condições não terão nenhuma previsão associada e não entrarão na medição do erro.

Para cada uma das 100 amostras, os erros de previsão para o número de transações, o valor médio da transação por cliente e o CLV estão representados na Tabela 22.

Tabela 22 RMSE e MAE e respetivos desvios-padrão, e o coeficiente de correlação de Spearman do Modelo de Pareto/NBD para as 100 amostras estratificadas

amos	error CLV		error_cesta		error_trx		square_error CLV		Spear	amos	error CLV		error_cesta		error_trx		square_error CLV		Spear
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	man		Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	man
0	332,8	850,5	4,1	4,2	8,8	19,5	912,9	3 788,1	0,85	**51**	313,3	407,0	4,2	5,1	7,3	13,1	513,4	998,7	0,83
1	314,7	385,5	4,7	6,6	9,6	35,9	497,5	883,8	0,85	**52**	311,1	368,9	4,4	4,9	8,7	20,9	482,4	896,1	0,85
2	317,2	428,4	4,3	5,6	7,6	15,4	532,9	1 184,9	0,86	**53**	326,9	449,8	4,4	4,8	7,0	11,4	555,9	1 218,2	0,83
3	292,5	363,2	4,5	4,8	7,1	14,9	466,2	865,9	0,87	**54**	319,4	518,1	4,5	5,6	8,6	16,2	608,4	1 997,8	0,84
4	283,8	325,9	4,8	7,7	9,0	20,8	432,0	718,5	0,86	**55**	314,4	407,4	4,2	4,6	7,3	14,2	514,5	986,7	0,86
5	313,9	398,6	4,5	4,6	8,0	14,9	507,2	902,9	0,85	**56**	336,9	480,7	3,9	4,2	9,2	37,0	586,8	1 217,3	0,84
6	301,0	417,7	4,6	6,0	7,7	17,3	514,7	1 248,1	0,87	**57**	300,6	338,1	4,4	4,9	7,7	17,6	452,3	786,0	0,85
7	312,9	351,5	5,0	6,3	7,7	16,3	470,4	822,7	0,84	**58**	312,9	422,9	4,3	5,4	8,6	15,3	525,9	1 152,3	0,84
8	293,2	405,1	4,4	5,4	7,3	12,3	499,9	1 063,2	0,84	**59**	302,6	340,8	4,9	5,8	7,3	12,7	455,7	777,4	0,85
9	318,2	411,3	4,1	4,8	8,5	18,2	519,8	993,5	0,84	**60**	298,5	419,7	4,4	4,9	7,2	15,6	514,8	1 174,3	0,87
10	286,1	364,1	4,5	5,9	8,1	15,6	462,9	893,4	0,85	**61**	321,7	443,9	4,9	17,4	8,7	21,2	548,0	1 118,4	0,84
11	300,0	381,3	5,0	5,8	9,4	23,6	485,0	998,6	0,88	**62**	287,7	363,9	4,1	4,5	8,1	15,9	463,8	794,5	0,85
12	314,6	429,3	4,4	5,1	8,1	16,2	532,1	1 048,9	0,85	**63**	307,3	348,4	4,5	5,0	8,4	15,2	464,5	732,5	0,83
13	298,9	343,3	4,5	5,1	8,3	17,0	455,0	754,9	0,86	**64**	303,4	349,4	4,2	4,7	8,1	19,2	462,6	757,6	0,84
14	317,3	392,7	4,2	4,1	8,1	12,7	504,7	1 085,5	0,85	**65**	333,8	702,2	3,9	4,5	8,0	15,0	777,2	3 347,7	0,85
15	287,2	408,9	4,1	5,0	8,4	18,3	499,5	1 089,0	0,85	**66**	312,6	442,5	4,1	4,5	8,0	15,8	541,6	1 128,4	0,87
16	301,9	352,5	4,9	5,6	7,7	14,4	464,0	768,2	0,84	**67**	309,5	381,3	4,6	5,0	7,6	12,8	490,9	990,7	0,84
17	326,2	470,6	4,6	5,4	8,1	17,5	572,4	1 503,9	0,83	**68**	294,0	344,7	3,9	4,0	7,5	14,2	452,9	784,6	0,84
18	304,7	511,3	4,4	5,2	8,0	14,5	595,0	2 108,6	0,84	**69**	316,2	447,0	4,2	4,9	8,9	17,8	547,4	1 464,5	0,84
20	299,9	358,7	4,3	5,1	7,2	14,5	467,4	809,0	0,86	**70**	316,1	442,6	4,4	7,3	7,8	17,6	543,7	1 392,5	0,83
21	303,0	458,4	4,9	6,7	8,8	19,3	549,3	1 704,7	0,83	**71**	299,0	368,6	4,7	7,9	8,1	14,8	474,5	914,1	0,85
22	322,5	441,7	4,5	7,2	7,3	13,9	546,7	1 124,0	0,84	**72**	322,4	382,3	4,4	4,8	7,0	12,9	499,9	823,0	0,83
23	308,5	444,7	4,3	4,6	8,6	19,5	541,1	1 285,3	0,85	**73**	307,6	405,4	4,5	4,8	8,4	15,9	508,7	985,0	0,84
24	298,8	337,2	4,3	5,0	7,9	14,4	450,4	717,5	0,84	**74**	328,7	442,4	4,7	6,2	8,7	17,8	550,9	1 131,3	0,84
25	311,6	421,2	4,7	6,3	9,3	28,8	523,7	1 079,8	0,84	**75**	292,1	344,3	4,2	4,5	7,7	16,7	451,4	789,0	0,86
26	311,6	376,2	4,4	4,9	6,8	12,3	488,3	821,1	0,84	**76**	304,8	376,0	4,6	5,2	9,8	29,6	483,9	1 066,8	0,84
27	297,8	383,5	4,3	5,3	8,2	15,9	485,3	887,0	0,85	**77**	293,6	358,4	4,3	5,1	7,7	17,4	463,1	795,5	0,85
28	324,4	412,3	4,3	5,1	8,9	17,0	524,5	980,8	0,83	**78**	283,8	330,8	4,7	5,1	7,0	12,5	435,8	808,1	0,85
29	315,7	399,1	4,3	4,4	8,1	20,2	508,7	1 008,9	0,86	**79**	315,1	399,9	4,1	5,7	8,0	13,7	508,9	992,7	0,82
30	298,3	359,8	4,1	5,3	8,6	17,2	467,2	845,2	0,85	**80**	336,0	455,4	4,6	5,2	7,5	13,5	565,7	1 114,7	0,84
31	294,7	346,1	4,8	8,1	8,5	18,6	454,4	770,0	0,86	**81**	297,8	381,3	4,4	5,3	8,3	14,5	483,6	949,1	0,84
32	298,4	380,5	4,6	6,9	7,5	13,0	483,4	914,2	0,83	**82**	318,0	394,6	4,2	6,6	8,1	16,7	506,6	901,6	0,85
33	317,6	461,8	4,6	5,2	9,1	20,1	560,3	1 229,0	0,85	**83**	292,8	341,4	4,0	6,9	8,9	17,7	449,6	767,2	0,83
34	312,1	395,0	3,9	4,2	7,5	13,7	503,3	961,4	0,83	**84**	313,6	382,6	4,2	4,6	7,5	21,0	494,5	939,7	0,83
35	337,1	540,6	4,4	6,7	8,4	17,4	636,9	1 753,8	0,83	**85**	311,5	386,5	4,7	5,4	7,6	13,1	496,2	951,5	0,85
36	309,5	390,8	4,5	5,5	8,7	16,1	498,4	1 034,7	0,84	**86**	306,3	400,3	4,1	4,5	7,8	14,6	503,9	928,0	0,84
37	301,0	402,4	4,4	5,1	8,4	18,8	502,3	1 076,1	0,84	**87**	292,1	377,9	4,7	11,4	8,5	18,5	477,5	896,3	0,86
38	286,9	341,4	4,2	4,6	8,0	16,0	445,8	783,3	0,85	**88**	315,0	412,8	4,2	4,7	8,6	16,6	519,1	983,1	0,84
39	299,8	375,1	4,5	5,1	7,7	13,3	480,1	881,3	0,83	**90**	326,1	517,6	4,2	4,7	7,6	14,2	611,5	1 756,0	0,84
40	315,7	409,1	4,5	6,0	8,6	15,6	516,6	971,6	0,86	**91**	319,3	398,6	4,7	6,4	7,8	16,2	510,5	928,5	0,85
41	318,1	405,5	4,7	7,1	7,9	17,4	515,2	966,7	0,84	**92**	303,3	390,6	4,8	6,6	7,9	15,9	494,4	1 086,1	0,84
42	308,0	361,7	4,2	6,6	8,3	17,7	475,0	868,4	0,84	**93**	329,5	392,8	4,7	4,8	8,4	26,2	512,5	878,3	0,83
43	318,2	385,3	4,4	4,9	7,7	14,5	499,5	904,5	0,84	**94**	321,9	391,8	4,6	5,7	8,4	14,0	506,9	886,2	0,83
44	310,5	456,8	4,1	4,1	7,5	15,4	552,1	1 348,8	0,85	**95**	312,2	399,9	4,8	8,0	7,3	12,7	507,2	972,1	0,86
45	326,9	675,4	4,6	5,8	9,0	16,2	750,0	2 724,5	0,87	**96**	309,4	363,0	4,8	6,5	7,5	15,1	476,8	839,2	0,85
46	299,5	363,5	4,4	6,2	8,9	24,4	470,8	839,2	0,83	**97**	296,2	368,6	4,0	4,4	8,3	14,5	472,7	853,6	0,84
47	302,3	359,6	4,1	5,0	7,3	12,8	469,6	887,0	0,87	**99**	304,8	459,1	4,5	7,5	7,0	13,3	550,9	1 608,4	0,82
48	319,0	459,5	4,3	5,1	7,7	17,1	559,2	1 218,3	0,85	**100**	313,4	402,7	4,2	4,5	8,1	13,4	510,1	994,9	0,84
49	303,7	359,0	4,5	6,3	8,3	19,2	470,1	829,4	0,86	**101**	331,4	428,9	4,1	4,4	8,4	17,7	541,8	1 067,3	0,80
50	307,9	411,0	4,7	5,0	7,6	12,7	513,3	1 174,0	0,86	**102**	311,5	385,8	4,7	6,5	8,4	15,4	495,7	994,3	0,85

Os erros de previsão para o número de transações e valor gasto médio por transação são ligeiramente inferiores. Isto acontece porque estes valores em número absoluto são valores inferiores ao valor do CLV. Os valores de MAE aparentam ser ligeiramente inferiores aos registados pela regressão. Os coeficientes de *Spearman* apresentam

também correlações próximas de 80% o que indicia um bom ajustamento do Modelo de Pareto.

4.3 Regressão Linear com as variáveis de entrada do Modelo Pareto/NBD

O pensamento que está subjacente ao desenvolvimento da Regressão Linear com as variáveis de entrada do Modelo Pareto/NBD é perceber se se trata de um melhor/pior método para prever o CLV ou se são as variáveis de entrada utilizadas que influenciaram os resultados. Com esta abordagem tem-se um mesmo ponto de partida para o desenvolvimento das duas metodologias.

As variáveis utilizadas no Modelo Pareto/NBD para previsão do número de transações são o número de transações repetidas, a *recency* (t) e o *cohort* (T_cohort) e para a previsão posterior do gasto de cada cliente são utilizadas as VLR de cada transação.

Segundo Glady (2009), na regressão, em vez de se utilizar o número de transações, deve-se usar VLR total, uma vez que este é o produto do número de transações pelo valor de VLR por transação. Como já fora concluído pela análise da matriz de correlações

The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: vlr_resp

Number of Observations Read	3221248
Number of Observations Used	3221248

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	2.084712E12	1.042356E12	9282444	<.0001
Error	3.22E6	3.617241E11	112293		
Corrected Total	3.22E6	2.446436E12			

Root MSE	335.10187	R-Square	0.8521
Dependent Mean	735.94747	Adj R-Sq	0.8521
Coeff Var	45.53340		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	39.41971	0.46778	84.27	<.0001	0	.	0	38.50287	40.33655
VLR	VLR	1	0.93560	0.00024961	3748.18	<.0001	0.92230	0.75808	1.31913	0.93511	0.93609
t		1	0.01212	0.00181	6.70	<.0001	0.00165	0.75808	1.31913	0.00857	0.01567

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	VLR	t
1	2.54962	1.00000	0.02140	0.04953	0.01763
2	0.37739	2.59922	0.10949	0.78527	0.01911
3	0.07299	5.91018	0.86912	0.16520	0.96326

Figura 27 Regressão Linear com VLR e t – ANOVA, principais medidas de desempenho do modelo, parâmetros estimados e diagnóstico de multicolinearidade

apresentada no capítulo 3, as variáveis t e T_cohort estão correlacionadas e a variável selecionada foi a variável t. A regressão que prevê o valor do CLV com base nas variáveis explicativas VLR e t, tem o *output* representado na Figura 27.

O modelo apresenta uma variância explicada de 85,21% ($R_a^2 = 0{,}8021$), um valor de RMSE de 335,10 e as duas variáveis consideradas são importantes na previsão do CLV. Relativamente aos pressupostos, não há indícios de existência de multicolinearidade entre as variáveis (VIF<5 e *Condition index* <10). Mas na análise de resíduos verifica-se que estes não apresentam uma variância constante nem uma distribuição normal. Isto pode dever-se há existência de *outliers*, como se pode verificar na Figura 28 , que estejam a criar ruído na previsão do CLV.

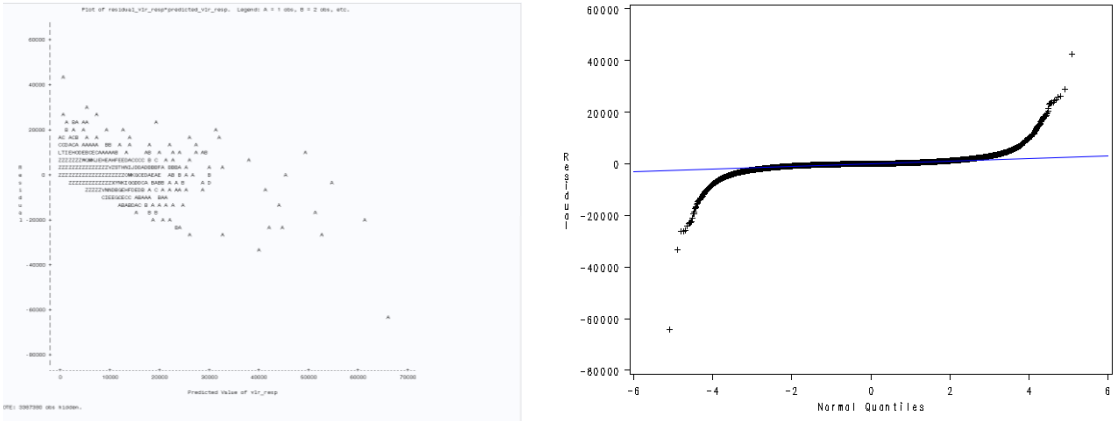


Figura 28 Regressão Linear com VLR e t - Diagrama de Dispersão e Gráfico Q-Q

Recorrendo à Distância de *Cook* excluem-se todas as observações que apresentem uma distância de *Cook* superior a 0,000001180. Assim excluíram-se 168 580 clientes e o diagrama de dispersão e o gráfico Q-Q sugerem uma distribuição normal, mas ainda uma variância não constante.

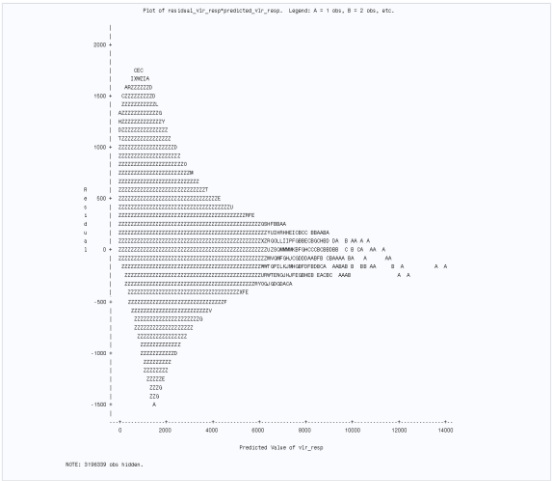


Figura 29 Regressão Linear com VLR e t - Diagrama de Dispersão dos resíduos versus o CLV estimado, após eliminação de outliers com distância de *Cook* > 0,000001180

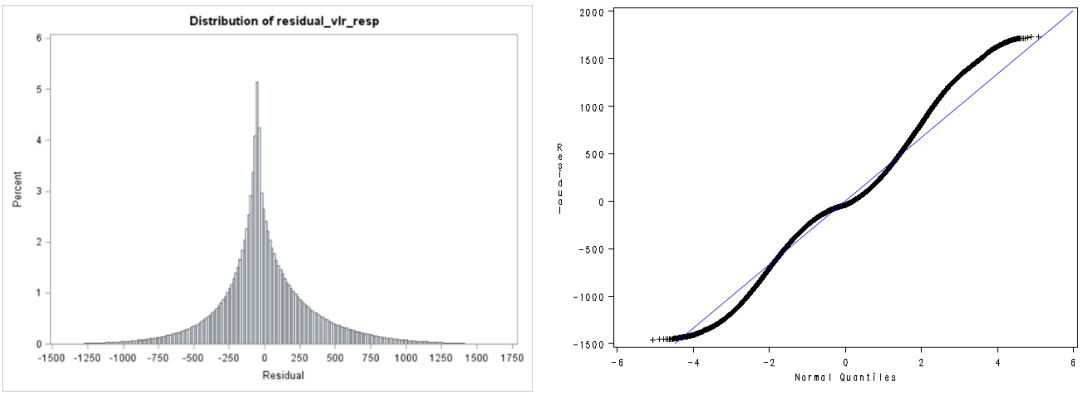


Figura 30 Regressão Linear com VLR e t – Histograma e gráfico Q-Q dos resíduos, após eliminação de outliers com distância de Cook > 0,000001180

Tal como foi executado no desenvolvimento da Regressão Linear para o conjunto total de observações, a transformação adequada a aplicar quando a variância dos resíduos não é constante ao longo dos diferentes valores previstos é a Logaritmização. Far-se-á aqui a mesma transformação considerada antes. O resultado está representado na Figura 31, onde é ainda possível verificar a existência de várias observações *outliers*.

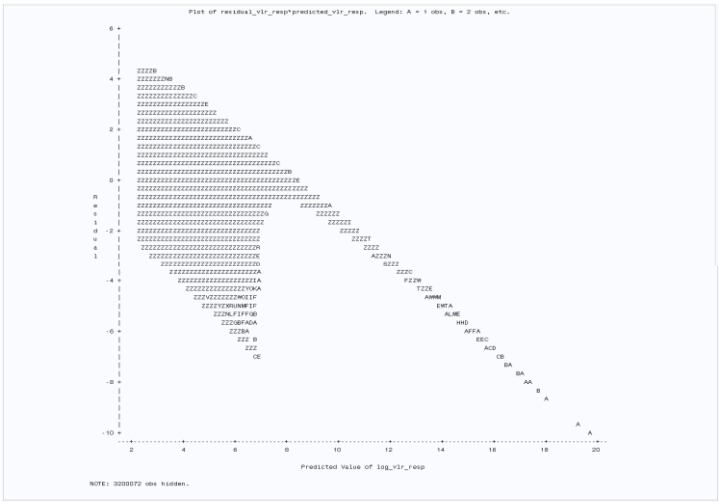
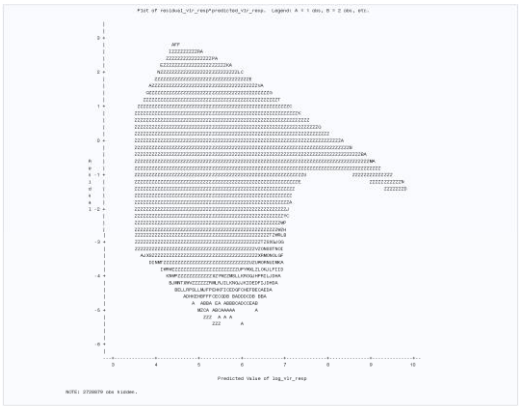
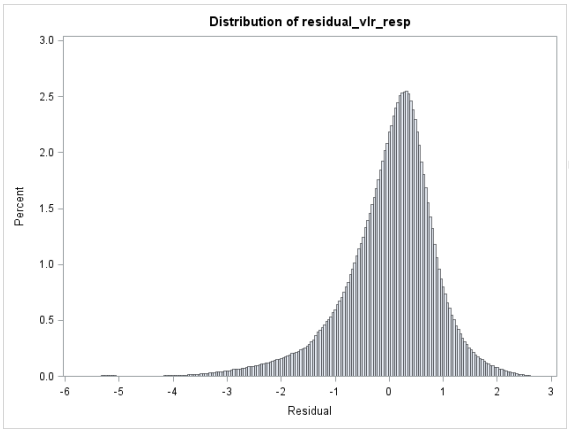
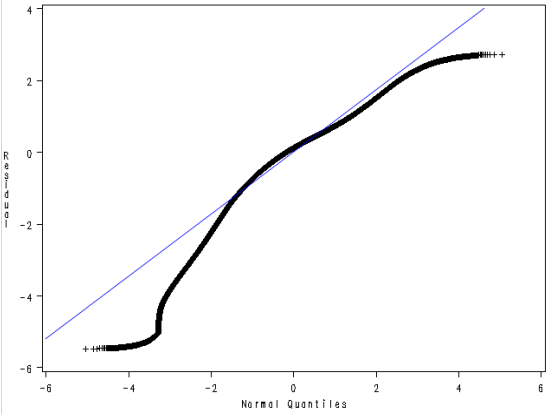


Figura 31 Regressão Linear com VLR e t – Diagrama de Dispersão dos resíduos versus o CLV estimado, após eliminação de outliers com distância de Cook > 0,000001180 e logaritmação da variável resposta

Com a exclusão das observações *outliers* o *output* final com o modelo encontra-se representado na Figura 32.







The REG Procedure

Model: Linear_Regression_Model

Dependent Variable: log_vlr_resp

Number of Observations Read	2775770
Number of Observations Used	2775770

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3422246	1711123	2278006	<.0001
Error	2.78E6	2085016	0.75115		
Corrected Total	2.78E6	5507262			

Root MSE	0.86669	R-Square	0.6214
Dependent Mean	5.99652	Adj R-Sq	0.6214
Coeff Var	14.45320		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	3.55448	0.00185	1919.38	<.0001	0	.	0	3.55085	3.55811
VLR	VLR	1	0.00098655	7.3187E-7	1347.99	<.0001	0.56289	0.78218	1.27848	0.00098512	0.00098798
t		1	0.00551	0.00000660	834.56	<.0001	0.34850	0.78218	1.27848	0.00550	0.00552

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation		
			Intercept	VLR	t
1	2.63060	1.00000	0.01038	0.04333	0.00878
2	0.33308	2.81030	0.05279	0.81415	0.01578
3	0.03632	8.51070	0.93683	0.14253	0.97544

Figura 32 *Output* completo do modelo de Regressão Linear final com as variáveis VLR e t

O modelo final cumpre todos os pressupostos para aplicação do modelo de Regressão Linear, explica 62,14% da variabilidade do CLV e tem um valor de RMSE no valor de 945,2607.

O modelo escreve-se então assim:

$$\widehat{vlr\_resp} = 3{,}554 + \ 0{,}00098VLR + 0{,}0055t$$

Tal como no modelo para o conjunto total a variável VLR apresenta uma contribuição positiva na previsão do CLV. A variável do intervalo de tempo entre a primeira e última compra também contribui positivamente para a previsão do CLV. Isto porque: quanto maior for o intervalo de tempo significa que o cliente apresenta compras no início e no fim do período de análise, sugerindo a sua permanência durante o período.

Validação com o método *Holdout*

Para uma correta validação do modelo dividiu-se o conjunto de dados na proporção de 70/30, sendo que a amostra de 70% foi selecionada através de amostragem estratificada, tal como efetuado na Regressão Linear com as 17 variáveis. Os erros estão representados na Tabela 23.

Tabela 23 RMSE e MAE e respetivos desvios-padrão para 99% das melhores previsões

m_mae	std_mae	m_rmse	std_rmse
2209,887	8939,212	9208,313	26242,608

segm_valor	m_mae	std_mae	m_rmse	std_rmse	nr_cli
Loyal	10.647,3	18.803,2	21.608,4	39.594,2	164.253
Frequent	810,9	3.556,6	3.647,8	15.871,7	383.208
Occasional	183,3	373,2	415,8	2.948,3	431.475
Sem Valor	3.095,7	10.777,4	11.213,0	29.248,8	27.843

Resultados para as 100 amostras

Da mesma forma dos restantes dois modelos, foi efetuada a previsão para as mesmas 100 amostras de clientes. Calcularam-se as três medidas de avaliação de desempenho e os resultados estão representados na Tabela 24. É importante referir que, tal como foi

efetuado nos modelos acima, as medidas de desempenho foram calculadas tendo por base 99% das melhores previsões.

Tabela 24 RMSE e MAE e respetivos desvios-padrão, e o coeficiente de correlação de Spearman do Modelo de Regressão com VLR e t para as 100 amostras estratificadas

amo	error		square_error		Spearman
	Mean	Std Dev	Mean	Std Dev	
0	710,2	2 445,1	2 545,0	7 431,8	0,83
1	627,9	1 828,2	1 932,1	5 350,2	0,82
2	536,4	1 399,7	1 498,3	4 148,8	0,83
3	493,5	1 196,3	1 293,6	3 908,5	0,84
4	533,0	1 498,5	1 589,8	4 500,1	0,82
5	633,4	1 836,3	1 941,6	5 046,7	0,82
6	746,9	2 952,1	3 043,7	8 952,6	0,83
7	560,5	1 323,2	1 436,4	3 346,6	0,81
8	516,5	1 304,5	1 402,4	3 834,8	0,81
9	622,8	1 884,1	1 983,5	5 448,8	0,82
10	596,3	1 881,8	1 973,2	5 826,4	0,80
11	864,2	4 386,4	4 468,5	15 047,1	0,85
12	675,2	2 054,6	2 161,7	5 870,6	0,82
13	645,2	2 214,3	2 305,4	6 987,6	0,83
14	670,3	2 300,2	2 394,7	8 399,6	0,82
15	555,8	1 519,9	1 617,6	4 500,8	0,80
16	582,9	1 830,4	1 920,1	5 590,5	0,82
17	546,4	1 593,7	1 684,0	4 908,9	0,81
18	436,4	1 102,4	1 185,1	3 456,8	0,82
20	537,0	1 582,2	1 670,1	5 410,9	0,82
21	571,8	1 990,7	2 070,2	6 963,1	0,78
22	936,6	3 448,4	3 571,6	10 288,0	0,80
23	771,5	3 145,4	3 237,1	10 248,2	0,81
24	518,4	1 585,7	1 667,6	5 160,5	0,82
25	624,5	2 042,3	2 134,7	6 575,9	0,85
26	482,2	1 227,8	1 318,5	3 737,2	0,81
27	890,4	4 326,2	4 414,7	15 177,6	0,82
28	628,1	1 811,8	1 916,7	5 789,8	0,83
29	614,6	1 879,5	1 976,6	5 329,2	0,85
30	509,6	1 501,4	1 584,8	4 752,6	0,83
31	595,6	1 787,4	1 883,1	5 165,2	0,85
32	515,3	1 378,5	1 471,0	4 264,4	0,81
33	619,5	1 800,7	1 903,4	5 155,2	0,81
34	515,2	1 448,9	1 537,1	4 511,8	0,82
35	953,9	3 899,0	4 012,1	12 059,2	0,81
36	602,1	1 738,1	1 838,6	5 159,9	0,82
37	546,4	1 464,2	1 562,1	4 086,0	0,82
38	436,1	942,7	1 038,2	2 492,1	0,83
39	519,7	1 418,1	1 509,6	4 277,9	0,81
40	641,8	2 252,9	2 341,5	7 216,1	0,83
41	667,0	2 557,1	2 641,4	8 672,9	0,81
42	483,6	1 310,3	1 396,1	4 043,6	0,80
43	595,9	1 678,8	1 780,6	4 731,3	0,81
44	602,6	1 983,0	2 071,5	6 156,8	0,81
45	1 063,3	6 141,8	6 230,1	24 604,4	0,83
46	485,4	1 164,6	1 261,2	3 584,4	0,80
47	568,1	1 510,1	1 612,7	4 217,1	0,84
48	652,3	2 333,3	2 421,6	7 673,1	0,83
49	657,3	2 222,8	2 316,9	6 934,1	0,82
50	539,3	1 441,3	1 538,2	3 966,7	0,82

amo	error		square_error		Spearman
	Mean	Std Dev	Mean	Std Dev	
51	516,4	1 239,8	1 342,5	3 636,4	0,81
52	722,0	2 637,6	2 733,4	8 262,3	0,83
53	461,0	1 356,8	1 432,3	4 523,2	0,81
54	676,5	2 523,4	2 611,3	8 380,3	0,81
55	599,9	1 719,8	1 820,6	5 232,0	0,84
56	474,2	1 003,4	1 109,4	2 741,4	0,83
57	529,4	1 213,4	1 323,3	3 863,9	0,82
58	596,3	1 947,6	2 035,9	6 349,9	0,82
59	479,3	1 169,6	1 263,4	3 266,5	0,82
60	658,7	2 091,8	2 192,0	6 137,9	0,84
61	607,0	1 946,2	2 037,7	6 056,2	0,83
62	563,5	1 501,5	1 603,0	4 097,6	0,83
63	673,8	2 287,9	2 384,0	7 436,2	0,80
64	594,5	1 961,9	2 049,0	6 361,3	0,81
65	785,2	2 695,9	2 806,6	7 935,8	0,82
66	956,2	4 889,5	4 979,7	16 035,0	0,83
67	494,7	1 355,0	1 441,8	4 601,4	0,81
68	432,7	862,4	964,5	2 248,8	0,82
69	555,9	1 447,2	1 549,6	4 300,9	0,82
70	487,7	1 151,1	1 249,6	3 373,9	0,80
71	604,5	1 882,1	1 975,9	5 763,8	0,83
72	551,1	1 401,4	1 505,3	4 077,0	0,83
73	568,7	1 505,0	1 608,2	4 270,2	0,82
74	656,5	2 158,7	2 255,3	6 604,5	0,82
75	615,2	1 809,3	1 910,2	5 376,2	0,83
76	659,0	2 113,2	2 212,6	6 209,4	0,83
77	459,7	994,7	1 095,3	2 804,3	0,83
78	668,8	2 619,0	2 701,7	8 728,0	0,84
79	515,5	1 400,8	1 491,9	4 250,7	0,79
80	594,3	1 509,9	1 622,0	4 089,2	0,84
81	521,3	1 421,8	1 513,7	4 032,4	0,82
82	536,4	1 327,0	1 430,6	3 835,9	0,81
83	511,8	1 206,7	1 310,2	3 643,2	0,80
84	569,1	1 713,2	1 804,5	5 417,1	0,82
85	514,7	1 400,0	1 491,0	4 348,8	0,82
86	688,0	2 615,3	2 703,0	8 914,9	0,82
87	539,6	1 491,7	1 585,6	4 456,5	0,82
88	789,2	3 887,1	3 964,5	14 610,7	0,81
90	590,8	1 676,4	1 776,6	5 085,5	0,81
91	892,3	3 600,0	3 707,2	11 180,0	0,83
92	524,1	1 278,0	1 380,7	3 720,6	0,82
93	733,5	2 841,5	2 933,3	10 250,0	0,79
94	524,2	1 284,6	1 386,9	3 688,1	0,80
95	550,7	1 429,3	1 531,0	4 260,5	0,83
96	678,7	2 471,4	2 561,8	8 143,4	0,83
97	440,6	970,1	1 065,0	2 824,4	0,81
99	444,9	1 021,0	1 113,2	3 010,4	0,82
100	548,5	1 450,0	1 549,6	4 444,7	0,82
101	571,5	1 605,5	1 703,4	4 781,4	0,82
102	623,3	1 868,8	1 969,1	6 036,3	0,82

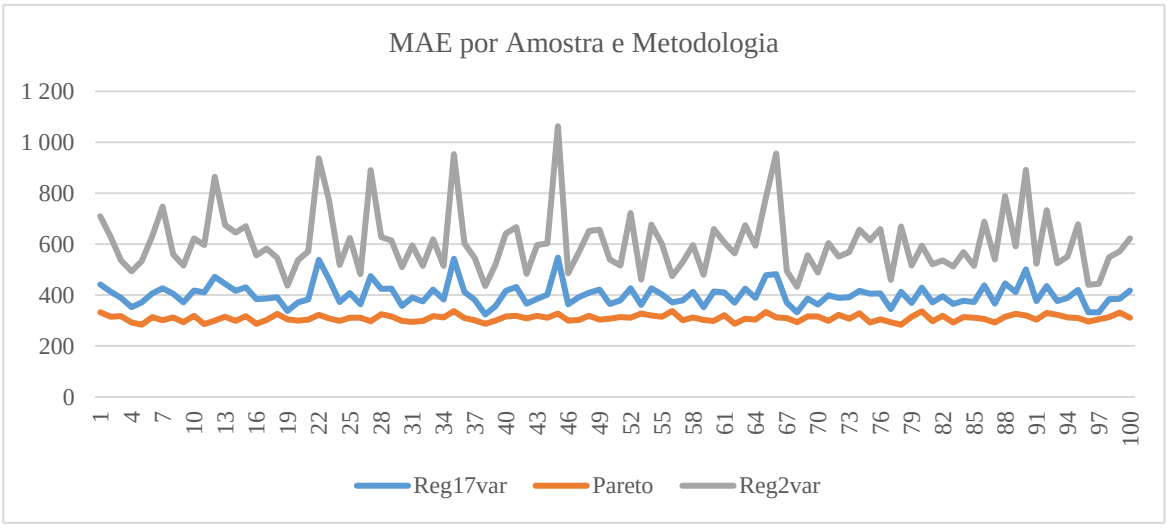
5 Discussão e Análise de Resultados

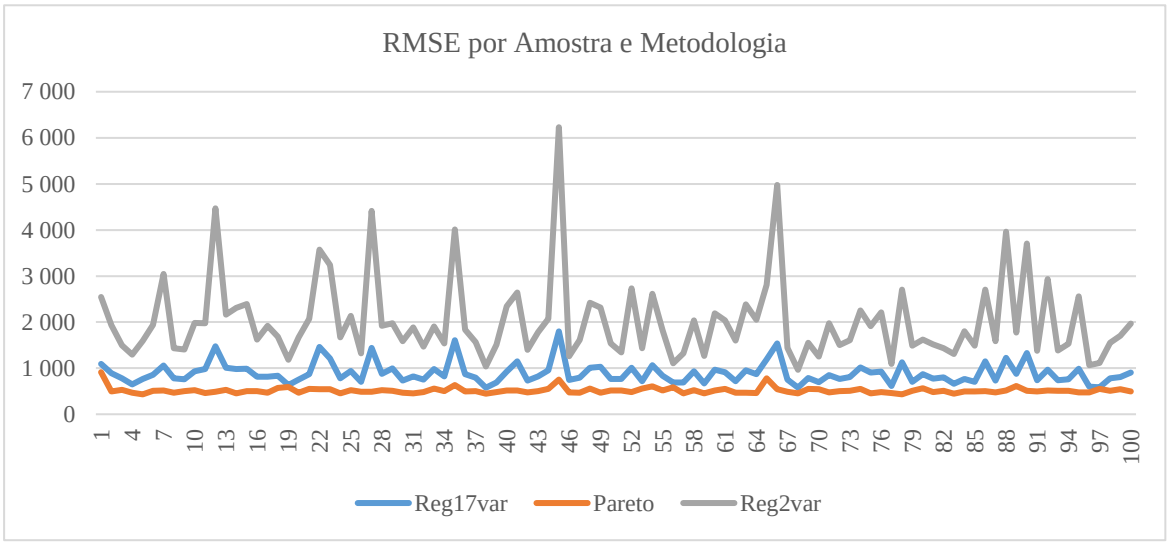
Nesta dissertação foram desenvolvidos três modelos, tendo por base duas metodologias distintas, com o objetivo de perceber qual será a melhor abordagem para a prever o CLV do cliente no período de um ano. As medidas que irão sustentar esta comparação serão o RMSE, MAE (com 99% das melhores previsões) e o coeficiente de correlação de *Spearman* entre a variável resposta e a respetiva previsão, sendo assim possível medir a assertividade do modelo em valor e em grandeza.

As vantagens já conhecidas dos dois modelos focam-se essencialmente na liberdade de desenvolvimento e na robustez de cada modelo. A regressão permite a consideração de diferentes variáveis, que dependendo do negócio, podem ser essenciais na previsão do CLV. Este caso é impossível no Modelo Pareto/NBD, quando se recorre ao *package* BTYD porque é um modelo com as variáveis de entrada bem definidas. Mas por outro lado o Modelo de Pareto/NBD é mais robusto, sem necessidade de dividir o conjunto total de dados para estimação dos parâmetros, tal como acontece no Modelo de Regressão Linear.

Tendo em conta as limitações inerentes à dimensão da amostra a comparação será efetuada sobre os erros avaliados nas 100 amostras estratificadas. Na Figura 33 estão representadas as três medidas de desempenho para os três modelos desenvolvidos. Para mais fácil representação dos modelos criou-se uma abreviação: Reg17var para a Regressão Linear composta pelas 17 variáveis de negócio, Pareto para o Modelo de Pareto/NBD e Reg2var para a Regressão Linear com a VLR e t (*recency*).

É possível concluir claramente que ao nível dos valores de MAE e RMSE o **Modelo de Pareto/NBD** é o que possui melhores previsões. É também o modelo de Pareto que possui pouca variabilidade de valores de MAE e RMSE ao longo das diferentes amostras. Nos valores do coeficiente de *Spearman* a conclusão não é assim tão clara. Os modelos de Pareto/NBD e Regressão Linear com as dezassete variáveis de negócio apresentam valores semelhantes, havendo amostras onde o modelo de Pareto apresenta melhores resultados e noutras o modelo Reg17var. O Modelo de Regressão Linear com as 2 variáveis de entrada do Modelo de Pareto/NBD é o que apresenta uma maior variabilidade nos erros quando





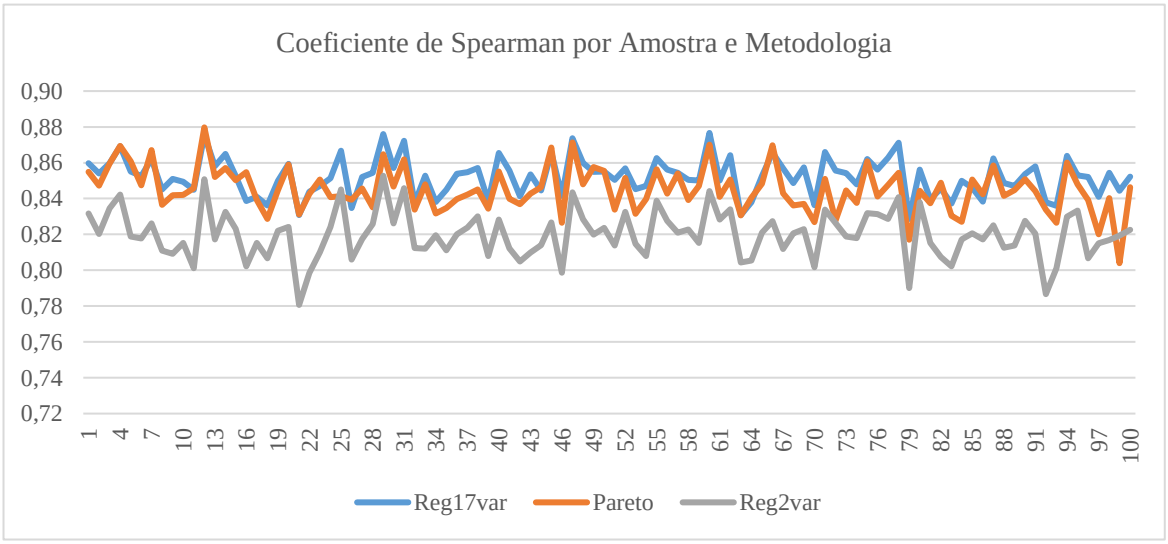


Figura 33 Medidas de desempenho para a Regressão Linear com as 17 variáveis, a Regressão Linear com as 2 variáveis e o Pareto/NBD

calculados em diferentes amostras, demonstrando que se trata de um modelo pouco robusto.

O Reg17var apresenta melhores previsões do que com o Reg2var, tanto no conjunto total de observações como nas previsões nas 100 amostras. Isto confirma que a inserção de variáveis de negócio acrescenta valor à previsão.

6 Conclusão

Este trabalho teve como principal objetivo prever CLV por cliente de uma empresa de retalho alimentar portuguesa. Dado ao leque de metodologias propostas na literatura (Singh & Jain, 2013), foram desenvolvidos três modelos de previsão para a identificação de qual a melhor metodologia a ser aplicada no sentido de obter previsões mais assertivas.

O modelo de Pareto/NBD é o modelo que apresenta melhores previsões do valor do cliente a longo prazo, ao nível de todas as medidas de desempenho consideradas, e é também o que apresenta uma maior robustez, não alterando a sua precisão nos diferentes conjuntos de dados. Seguido do Modelo do Pareto/NBD está a Regressão Linear com as 17 variáveis, e por fim a Regressão com as variáveis VLR e *recency*. A Regressão Linear com as 17 variáveis ainda assim aproxima-se mais dos resultados reais do que a Regressão com as duas variáveis utilizadas no Modelo de Pareto/NBD.

A área de retalho alimentar, apesar de ser um negócio com elevada frequência e de necessidades básicas, não garante um padrão normal de comportamento por parte do cliente. Existem clientes que são afetados por fatores externos que provocam uma mudança drástica no comportamento de compra, que nenhuma variável de negócio ou sociodemográfica consegue justificar. Estes factos são limitações naturais que dificilmente são possíveis de contornar na previsão. Além disso, a informação de comportamento do cliente na empresa, a adesão a ofertas da empresa e a sua informação sociodemográfica são fatores que explicam grande parte do valor que o cliente gastará nas lojas em estudo. Mesmo dentro da empresa, o acesso à informação ou até cruzamento de diferentes fontes não é simples, perdendo-se por vezes a oportunidade de obtenção de melhores resultados. Nesta dissertação, o investimento que a empresa efetua em cada cliente é uma variável essencial que, para além de melhorar os resultados obtidos, enriqueceria as conclusões retiradas e transmitiria, também, uma noção real do problema.

Para além das limitações no acesso a informação, foram encontradas também dificuldades no desenvolvimento dos modelos. A mais evidente é a dificuldade em trabalhar com conjunto de dados de grande dimensão. Os testes estatísticos, alguns *softwares*, e mesmo certas metodologias não são aplicáveis quando o conjunto de dados em estudo tem uma ordem de grandeza de milhões de observações. É o caso do Modelo do Pareto/NBD. O

software SAS não apresenta qualquer dificuldade no tratamento deste número elevado de observações, e por esta razão foi possível o desenvolvimento dos dois modelos de Regressão Linear para o conjunto total. Já o *software* R permite trabalhar com este volume de dados, mas com impacto na velocidade de processamento.

Para além de todo o trabalho desenvolvido ainda há abertura para progressos futuros ou versões secundárias dos modelos aqui descritos. A previsão do CLV, durante todo o seu processo, teve em conta os diferentes segmentos valor já existentes na empresa. Esta variável categórica garantiu que diferentes tipos de clientes estavam a ser analisados nas diferentes etapas dos modelos, na amostragem estratificada e avaliação do erro. Uma das hipóteses a serem consideradas no futuro é incluir esta variável na previsão do CLV. Outra hipótese seria desenvolver modelos preditivos distintos para os diferentes segmentos. Nesta abordagem, as variáveis a serem consideradas para cada modelo deveriam ter em conta o comportamento característico de cada segmento.

Seria também interessante a adição de uma variável que caracterizasse a evolução do cliente no período: se apresenta uma tendência de decréscimo ou aumento do seu envolvimento com a empresa. Provavelmente, um cliente que tendencialmente está a decrescer as suas vendas, a sua frequência ou outra medida de negócio, significa que no futuro terá um menor CLV do que o que seria esperado com um comportamento constante.

Contudo, cada vez é mais importante conhecer os clientes, quem são, quais são as preferências ou até mesmo as suas expectativas em relação à empresa, para que deste modo sejam identificados os fatores relevantes do passado que justificam as escolhas do presente e que sustentam a visão para o futuro.

Referências

Aghaie, A. (2009). Measuring and Predicting Customer Lifetime Value in Customer Loyalty Analysis: A Knowledge Management Perspective (A Case Study on an e-Retailer). *International Journal of Industrial Engineering & Production Research*, 20(1), 21-30.

Christensen, L. A. (1997). Introduction to building a linear regression model. In *Proceedings of the Twenty-Second Annual SAS Users Group International Conference.*

Dziurzynski, L., Wadsworth, E., & McCarthy, D. (2015). Package "BTYD." *CRAN Project*, 109. Retrieved from https://cran.r-project.org/web/packages/BTYD/BTYD.pdf

Eide, V. (2016). "125 Years" --Scientists Say Maximum Age for Humans Has Been Reached. Retrieved from http://www.dailygalaxy.com/my_weblog/2016/10/125-years-scientists-say-maximum-age-for-humans-has-been-reached.html

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005a). Counting Your Customers? the Easy Way: An Alternative to the Pareto/NBD Model. *Marketing Science*, *24*(2), 275–284.

Fader, P. S., Hardie, B. G., & Lee, K. L. (2005b). RFM and CLV: Using iso-value curves for customer base analysis. *Journal of Marketing Research*, 42(4), 415-430.

Gama, J., Carvalho, A. P. D. L., Faceli, K., Lorena, A. C., & Oliveira, M. (2015). *Extração de Conhecimento de Dados - Data Mining* (2ª edição). Lisboa: Edições Sílabo.

Glady, N., Baesens, B., & Croux, C. (2009). A modified Pareto/NBD approach for predicting customer lifetime value. *Expert Systems with Applications*, *36*(2), 2062–2071.

Gupta, S., Lehmann, D. R., & Stuart, J. A. (2004). Valuing Customers. *Journal of Marketing Research*, *41*(1), 7–18. http://doi.org/10.1509/jmkr.41.1.7.25084

INE. (2017). Famílias clássicas por número de indivíduos segundo os Censos - Portugal. Retrieved from

http://www.pordata.pt/Portugal/Famílias+clássicas+por+número+de+indivíduos+segundo+os+Censos-786

Khajvand, M., & Tarokh, M. J. (2011). Estimating customer future value of different customer segments based on adapted RFM model in retail banking context. *Procedia Computer Science*, *3*, 1327–1332.

Lin, M., Lucas Jr, H. C., & Shmueli, G. (2013). Research commentary—too big to fail: large samples and the p-value problem. *Information Systems Research*, 24(4), 906-917. http://doi.org/10.1287/isre.2013.0480

Maroco, J. (2007). *Análise Estatística com a utilização do SPSS*. Lisboa: Edições Sílabo.

McCarthy, D., & Wadsworth, E. (2014). *Buy ' Til You Die - A Walkthrough*. Disponível com o pacote R "BTYD: Implementing Buy 'Til You Die Models". https://cran.r-project.org/web/packages/BTYD/vignettes/BTYD-walkthrough.pdf

Petrini, J., Zulini, T. M., & Dias, R. A. P. (1999). Diagnóstico de multicolinearidade em modelos aditivo-dominante em uma população de bovinos de corte compostos (Bos taurus x Bos indicus), 2005–2007.

Pfeifer, P. E., Haskins, M. E., & Conroy, R. M. (2004). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. *Journal of Managerial Issues*, *8*(2), 239–258.

Schmittlein, D. C. ., Morrison, D. G. ., & Colombo, R. (1987). Counting Your Customers : Who Are They and What Will They Do Next ? *Management Science*, *33*(1), 1–24.

Singh, S., & Jain, D. (2013). Measuring Customer Lifetime Value: Models and Analysis. *INSEAD Working Papers Collection*, (27), 1–48.

Veroneze, R. (2011). *Tratamento de Dados Faltantes Empregando Biclusterização com Imputação Múltipla*. PhD thesis, Universidade Estadual de Canpinas.

Anexo 1. Análise da Multicolinearidade da Regressão Linear com o total de variáveis excluindo as correlacionadas e a variável perc_sem

Collinearity Diagnostics																					
			Proportion of Variation																		
Number	Eigenvalue	Condition Index	Intercept	VLR	DESCONTOS	t_recency	cesta_rep	primav	segm_baby_junior	segm_grocer	cart_uni	DL_SP	vb_eco	nr_ins	vb_nalim	FLAG_CLOSER_PARTNER	perc_vb_eco	idade_Median	agregado_Median	lifetime_Median	osd_Median
1	8,74205	1,000	0,00041	0,00132	0,00159	0,00111	0,00292	0,00024	0,00272	1,6E-05	0,00084	0,00196	0,00215	0,00232	0,00217	0,0019	0,0018	0,00083	0,00282	0,00161	0,00123
2	1,83429	2,183	0,00095	0,00915	0,014	0,08007	0,00055	0,00862	0,00762	0,00056	0,00013	0,01697	0,00654	0,00029	0,01998	0,00038	0,01273	0,00233	0,00703	0,00145	0,00298
3	1,22835	2,668	0,00013	0,00277	0,00407	0,00642	0,00852	0,02324	0,04212	0,00798	0,00916	0,00861	0,1409	0,01508	0,00113	0,01983	0,12077	0,00094	0,00151	0,00055	0,00065
4	1,02718	2,917	1,3E-05	3,5E-06	4,9E-05	8E-05	0,00018	0,35571	0,00768	0,01484	0,40428	0,01143	0,0108	0,00087	7,4E-05	0,00011	0,00957	5,3E-05	7,9E-05	6,5E-05	0,00013
5	1,00067	2,956	1,3E-05	7,1E-05	0,0002	0,00128	0,00011	0,00048	0,00268	0,94291	0,00951	3,9E-05	0,00344	1,9E-05	0,00047	0,00937	0,00146	7,4E-05	0,00019	0,00019	6,2E-05
6	0,92653	3,072	9,3E-06	2,3E-05	0,00041	0,02078	9E-05	0,58308	0,00304	0,00309	0,31931	0,01075	1,7E-06	0,00036	0,00083	0,00047	4,8E-05	3,4E-06	9,6E-05	8,43E-01	2,6E-05
7	0,84426	3,218	1,1E-05	0,00023	0,00267	0,02624	0,01365	0,00194	0,00076	0,01149	0,00269	0,00119	0,00196	0,00068	0,00308	0,85862	0,00099	8,5E-05	5,8E-05	2,1E-05	2,3E-05
8	0,63593	3,708	0,00018	0,00428	0,00614	0,04212	0,0003	0,00685	0,56733	3,6E-06	0,00681	0,00403	0,24455	0,00497	0,00114	0,00786	0,01151	7,09E-01	0,00522	0,00103	0,00068
9	0,56791	3,923	0,00077	0,00036	0,00197	0,55788	0,00193	0,00153	0,1719	0,00785	0,00633	0,00401	0,01847	0,01502	0,04139	0,04481	0,01386	0,00355	0,02905	0,00854	0,00377
10	0,41597	4,584	0,00024	0,0018	0,01474	0,06318	0,12182	0,00248	0,03179	0,00015	0,01161	0,03976	0,45368	0,03413	0,01517	0,00062	0,29922	0,00134	0,02995	0,00044	0,00066
11	0,36921	4,866	5,5E-05	0,00186	0,00123	0,05033	0,02588	0,00309	0,04248	0,00087	0,11532	0,32652	0,02738	0,00022	0,37523	0,00091	0,02819	0,00032	0,00056	2E-05	0,0002
12	0,34447	5,038	1,7E-05	0,00017	1,3E-05	0,01715	0,55056	0,00791	0,00032	0,00139	0,00263	0,00183	0,04175	0,00023	0,06339	0,01766	0,08479	1,9E-05	0,27341	0,00223	0,00188
13	0,30466	5,357	0,00261	0,00041	0,00113	0,00053	0,11692	0,00094	0,00037	1,2E-05	0,00664	0,0041	0,02897	0,00669	0,00057	0,01126	0,09163	0,01781	0,57658	0,04937	0,04205
14	0,22035	6,299	0,00038	0,00256	0,01059	0,0811	0,0198	0,00049	0,07049	0,00059	0,00823	0,03469	0,00091	0,81503	0,00436	3,98E-01	0,27351	0,01044	0,00078	0,0012	0,01463
15	0,16524	7,274	0,00097	0,00033	0,67928	0,00727	0,01846	0,0011	0,0142	0,00197	0,06491	0,29146	1,2E-05	0,00039	0,33923	9,4E-05	0,00031	0,00128	0,00309	0,0476	0,02845
16	0,14174	7,853	0,002	0,00947	0,07578	0,00372	0,00508	1,1E-05	0,00052	0,00138	0,00881	0,03731	4,1E-06	0,00929	0,05012	0,00068	0,00172	0,00095	2,8E-05	0,53579	0,32613
17	0,11851	8,589	0,0006	0,95895	0,17769	0,0032	0,02338	0,0001	0,00848	0,0042	0,02134	0,19897	0,01553	0,02989	0,07402	0,01699	0,03815	1,2E-05	1,8E-05	0,00503	0,00064
18	0,08435	10,180	0,01174	0,00096	0,00111	0,00404	0,04132	0,00089	0,01448	0,00016	5,6E-05	0,00049	5,4E-06	0,0116	0,00264	0,00102	0,00562	0,53127	0,00327	0,32385	0,31254
19	0,02831	17,571	0,9789	0,00527	0,00734	0,03352	0,04852	0,00128	0,01101	0,00055	0,00139	0,00587	0,00294	0,05293	0,005	0,00741	0,00412	0,4287	0,06626	0,02102	0,26327